Predictive analysis of auditory attention from physiological signals

Nikesh Bajaj

Supervisors:

Prof. Francesco Bellotti

Prof. Jesús Requena Carrión

Prof. Alessandro De Gloria

Joint Degree for Interactive and Cognitive Environment
University of Genova & Queen Mary University of London

This dissertation is submitted for the degree of
Doctor of Philosophy

JD-ICE April 2019

I dedicate this thesis to
my parents, my siblings, and my friends
for their constant support and unconditional love.
I love you all dearly.

Declaration

I hereby certify that I am the sole author of this thesis and that neither any part of this thesis nor whole of the thesis has been submitted for a degree or professional qualification to any other University or Institution

I certify that to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from work of other people included in my thesis, published or otherwise are fully acknowledged in accordance with the standard referencing practices.

Nikesh Bajaj
April 2019

Acknowledgements

The journey of my PhD has been a wonderful and a life-changing experience for me and it would not have been possible to complete this journey without the support and guidance that I received from many people.

First and foremost I want to thank my advisors Prof. Francesco Bellotti and Prof. Jesús Requena Carrión for the continuous support of my PhD study and related research, for their patience, motivation, productive and stimulating conversions, and immense knowledge. For pushing me to learn better and improve and a constant reminder of quality work. Their guidance helped me in all the time of research and writing of this thesis.

I also would like to express my sincere gratitude to Prof. Alessandro De Gloria and Prof. Riccardo Berta for their feedback and support towards writing articles and conducting the experiment.

Besides my advisor, I would like to thank my PhD program committee Prof. Carlo Regazzoni, Prof. Lucio Marcenaro, and Prof. Andrea Cavallaro, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives. My sincere thank also goes to Dr. Riccardo Mazzon for helping at various events, suggestions and including me in different activities. He has been a wonderful support. Very special gratitude goes out to all the staff members at Unige and QMUL for their support.

My time at Unige was made enjoyable in large part due to the many friends, labmates, and groups that became a part of my life. I am grateful for the time spent with my friends, long-lasting chats, and memorable trips. My time at QMUL was also enriched by the environment at the campus and the opportunities that I could avail by being there.

There are not enough words to thank my family. Special thank to my father for imparting in me with a lifestyle and attitude towards work, on which I have been climbing high. My mother Shanti Bajaj for her eternal love and support. I am grateful to my siblings, who not only supported me morally and emotionally but also gave me a push when I needed. I am also thankful to my younger siblings, Laxmi and Vipin, who always get ready to go with my crazy ideas. They have been wonderful friends to me. Thank you.

Abstract

In recent years, there has been considerable interest in recording physiological signals from the human body to investigate various responses. Attention is one of the key aspects that physiologists, neuroscientists, and engineers have been exploring. Many theories have been established on auditory and visual selective attention. To date, the number of studies investigating the physiological responses of the human body to auditory attention on natural speech is, surprisingly, very limited, and there is a lack of public datasets. Investigating such physiological responses can open the door to new opportunities, as auditory attention plays a key role in many cognitive functionalities, thus impacting on learning and general task performance.

In this thesis, we investigated auditory attention on the natural speech by processing physiological signals such as Electroencephalogram (EEG), Galvanic Skin Response (GSR), and Photoplethysmogram (PPG). An experiment was designed based on the well established dichotic listening task. In the experiment, we presented an audio stimulus under different auditory conditions: background noise level, length, and semanticity of the audio message. The experiment was conducted with 25 healthy, non-native speakers. The attention score was computed by counting the number of correctly identified words in the transcribed text response. All the physiological signals were labeled with their auditory condition and attention score. We formulated four predictive tasks exploiting the collected signals: Attention score, Noise level, Semanticity, and LWR (Listening, Writing, Resting, i.e., the state of the participant).

In the first part, we analysed all the user text responses collected in the experiment. The statistical analysis reveals a strong dependency of the attention level on the auditory conditions. By applying hierarchical clustering, we could identify the experimental conditions that have similar effects on attention score. Significantly, the effect of semanticity appeared to vanish under high background noise.

Then, analysing the signals, we found that the-state-of-the-art algorithms for artifact removal were inefficient for large datasets, as they require manual intervention. Thus, we introduced an EEG artifact removal algorithm with tuning parameters based on Wavelet Packet Decomposition (WPD). The proposed algorithm operates with two tuning parameters

and three modes of wavelet filtering: Elimination, Linear Attenuation, and Soft-thresholding. Evaluating the algorithm performance, we observed that it outperforms state-of-the-art algorithms based on Independent Component Analysis (ICA). The evaluation was based on the spectrum, correlation, and distribution of the signals along with the performance in predictive tasks. We also demonstrate that a proper tuning of the algorithm parameters allows achieving further better results.

After applying the artifact removal algorithm on EEG, we analysed the signals in terms of correlation of spectral bands of each electrode and attention score, semanticity, noise level, and state of the participant LWR). Next, we analyse the Event-Related Potential (ERP) on Listening, Writing and Resting segments of EEG signal, in addition to spectral analysis of GSR and PPG.

With this thesis, we release the collected experimental dataset in the public domain, in order for the scientific community to further investigate the various auditory processing phenomena and their relation with EEG, GSR and PPG responses. The dataset can be used also to improve predictive tasks or design novel Brain-Computer-Interface (BCI) systems based on auditory attention. We also use the deeplearning approach to exploit the spatial relationship of EEG electrodes and inter-subject dependency of a model. As a domain application, we finally discuss the implications of auditory attention assessment for serious games and propose a 3-dimensional difficulty model to design game levels and dynamically adapt the difficulty to the player status.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

APD Auditory Processing Disorder

BCI Brain Compute Interface

CNN Convolutional Neural Network

ECG Electrocardiography

EDA Electrodermal Activity

EEG Electroencephalogram

EMG Electromyogram

ERP Event Related Potential

fMRI Functional magnetic resonance imaging

fNIRS Functional Near-Infrared Spectroscopy

GSR Galvanic Skin Response

IBI Interbeat interval

ICA Independent Component Analysis

IPR Interpercentile range

IQR Interquartile-range

LSTM Long-Short Term Memory

LWR Listening, Writing, Resting

MAE Mean Absolute Error

MSE Mean Square Error

NPCs Non-player characters

PPG Photoplethysmogram

PSD Power Spectral Density

RNN Recurrent Neural Network

SG Serious Games

SNR Signal-to-noise ratio

SVM Support Vector Machine

WPD Wavelet Packet Decomposition

Chapter 1

Introduction

In this chapter, we will provide background information about the relative brain functionalities, well-established theories of cognitive psychology and measuring the brain activities using Electroencephalogram (EEG). This chapter also provides a brief detail of other physiological responses of the body. We end this chapter with the problem statement of the thesis and relevant state-of-the-art work.

1.1 The human brain

The human brain is perhaps the most complex systems that exist. Throughout history, many theories have been proposed to describe the brain structure and its functionalities. Some of them have been accepted by the wider audience by evaluating with scientific methods, yet many would say, it is not well understood. In order to provide the background information that is relevant to our work, we will describe the human brain with well established and widely accepted views.

The brain is considered to have three parts, the cerebrum, the cerebellum, and the brainstem. The cerebrum is the largest part of the human brain. The brainstem is situated at the beneath of the cerebrum, connected to the spinal cord. The cerebellum, also known as the little brain, a small structure located on the brainstem at the back of the cerebrum. The cerebrum is divided into two hemispheres; right hemisphere and left hemisphere. According to functionalities, each hemisphere is divided into four lobes, namely; the frontal, parietal, temporal and occipital lobes [1]. Combing four together, sometimes also called the cerebral cortex. The frontal lobe, as the name suggested is situated at the front part of the brain, towards the nose. The occipital lobe is located at the back side of the brain. The parietal is the middle and the temporal at sides of each hemisphere towards each ear. Both hemispheres are connected by mainly corpus callosum. Each hemisphere is being considered for different

functionality. The left brain is associated with a logical function such as mathematics, language. The right brain is associated with creativity such as imagination, expressions, art. Apart from this left-right categorization of function, different lobes are considered to be responsible for different functions. The frontal lobe is associated with executive functions for example planning, abstract thinking, reasoning, self-control etc. The frontal lobe is also responsible for the cognitive processes of the brain such as attention, memory, motivation. It contains the largest amount of dopamine neurons. The occipital lobe, as it is situated at the back of the brain is associated with visual information. The parietal lobe is dedicated to sensory information, mainly for skin such as touch, temperature, and pain. The temporal lobe is associated with the processing of visual and auditory information [1].

Depending on the field of interest, the brain has been studied in different areas. The medical doctors (Neurologist) study the brain to diagnose and treat brain diseases. The psychologist study brain to understand the behaviour of human being. Understanding the cognition processes such as attention, memory, and problem-solving are studied by the cognitive psychologist. All the cognitive functions such as filtering and processing the information, ability to retain information and think logically are processed by frontal lobe of the brain.

1.2 The cognitive attention

The attention is one of the most studied areas in cognitive psychology, including educational psychology and neuroscience. The attention is defined as a cognitive process of focusing on a discrete information while ignoring other, also known as selective attention. Two kinds of selective attention have been studied mainly; visual and auditory. However, some studies have investigated the task-oriented attention. The selective attention has been studied for a century to understand how brain selects one information to process over other [2] [3] and is a concept that has also been successfully applied to other sensory systems, such as the visual [4] [5]. Several theories have been proposed in the last century for explaining the mechanism of attention for both auditory and visual attention.

1.2.1 Auditory attention

Auditory attention became a topic of great interest in cognitive and neuroscience circles in the years following the Second World War, aroused by the documentation of cases where audible messages failed to be perceived by fighter pilots [6]. Two main theories have been proposed in the last century for explaining the mechanism of attention. The first one is known as filter

theory and suggests that the brain filters useful information over useless information [7]. Two opposing mechanisms have been proposed in the context of filter theory, namely early selection theory and late selection theory [8]. According to the early selection theory, the brain selects stimuli at early stages of information processing, whereas late selection theory suggests that the brain selects stimuli after semantic decoding only, i.e. at a later stage of information processing [9]. The second theory which provides a model for auditory attention is load theory and it gives a plausible unifying framework for early and late selection theory [10]. According to load theory, selection can be, at early or late stages, depending on the perceptual load of the stimuli [10]. Specifically, if the perceptual load of the stimuli is high, the brain will filter out useless information at early stages, whereas if the perceptual load of stimulus is low, then all the stimuli are processed before being filter out, which corresponds to a late selection process [9].

Most of the studies focusing on auditory attention involved simulated multi-message environments [11] and among them, the dichotic listening task has become one of the most popular settings. In the dichotic listening task, two different auditory stimuli are presented to the participants, who are asked to attend to only one of them [12]. After listening to the stimuli, participants are asked to write down the messages that were presented to them, which is used as an indication of the level of auditory attention [13].

1.2.2 Visual attention

The basic theories for selective attention are also applicable to visual attention, however visual attention is more complex than auditory attention. For selecting the visual information, we process the only fraction of the visual information from wide visual field available. The common experimental setting to investigate the visual attention is to record the reaction time of eye movement. The participant is presented with a stimulus either left or right side, with an indication where it likely to be with 80% of accuracy and measure the reaction time to correctly identify it [14]. Similar to dichotic listening task, in some setting, two superimposed videos are used in the experiment and asked to pay attention to the content of one video [15]. There are many different experiment setting for visual attention such as visual search, object-based tracking etc.

1.3 Measure of brain activity

The neural activity in the human brain is an electrical change. The brain process any information by means of neurons that use electrical and chemical signals to communicate by releasing

and receiving neurotransmitters. The electrical signals are generated in a brain throughout the life. Studying these electrical signals are vital to understanding the neurophysiological behaviour of the brain. A number of techniques are used to study brain activities. Functional magnetic resonance imaging (fMRI), Functional Near-Infrared Spectroscopy (fNIRS), and Electroencephalography (EEG) recordings widely used techniques. The fMRI measures the brain activity by scanning the blood flow. The fNIRS measures brain activity by measuring hemodynamic response in the brain through detecting the temporal changes in infrared light source. The EEG measures the electrical activity of the brain by electrodes placed on the scalp. Comparing to other two, EEG measures the brain activity directly, with high temporal resolution and most accessible and portable for the research. Since fMRI has a high spatial resolution but very expensive, it is mostly limited to medical diagnosis and treatments.

1.4 The EEG measure

The EEG signal is measured by placing electrodes on the scalp, that measure the current flow from neurons. Each neuron (brain cell), when activated, it produced an electrical and magnetic field around the scalp. The magnetic field is measured by Electromyogram (EMG), while an electric field is measured by EEG. Since there are 100 billion neurons in the brain, when an electrode is placed on the scalp, it measures the accumulative activity of many neurons together. The complex structure of the brain attenuates the electrical signals, therefore electrode can record the brain activity, only when a large number of neurons generate enough potential. The EEG devices amplify the recorded signal to store and process it.

The mounting literature of EEG studies has identified the five major frequency bands of the EEG signal and the placement of electrodes on the scalp. The frequency bands widely used are; Delta ($0.1-4$ Hz), Theta ($4-8$ Hz), Alpha ($8-14$ Hz), Beta ($14-30$ Hz), Gamma ($30-63$ Hz). A raw EEG signal and corresponding signal in different bands are shown in Figure 1.1. Figure 1.2 shows the topographic view of a scalp, with 14 EEG electrodes placement in 10-20 system. The 14-channel EEG signal and corresponding frequency bands with brain activity computed with the energy of the first second of each electrode are displayed with the color.

The frequency bands; Delta, Theta, Alpha, Beta, and Gamma, are also called as the brain rhythms. Brain waves have been investigated over decades and few characteristics behaviour of these brain waves have been established.

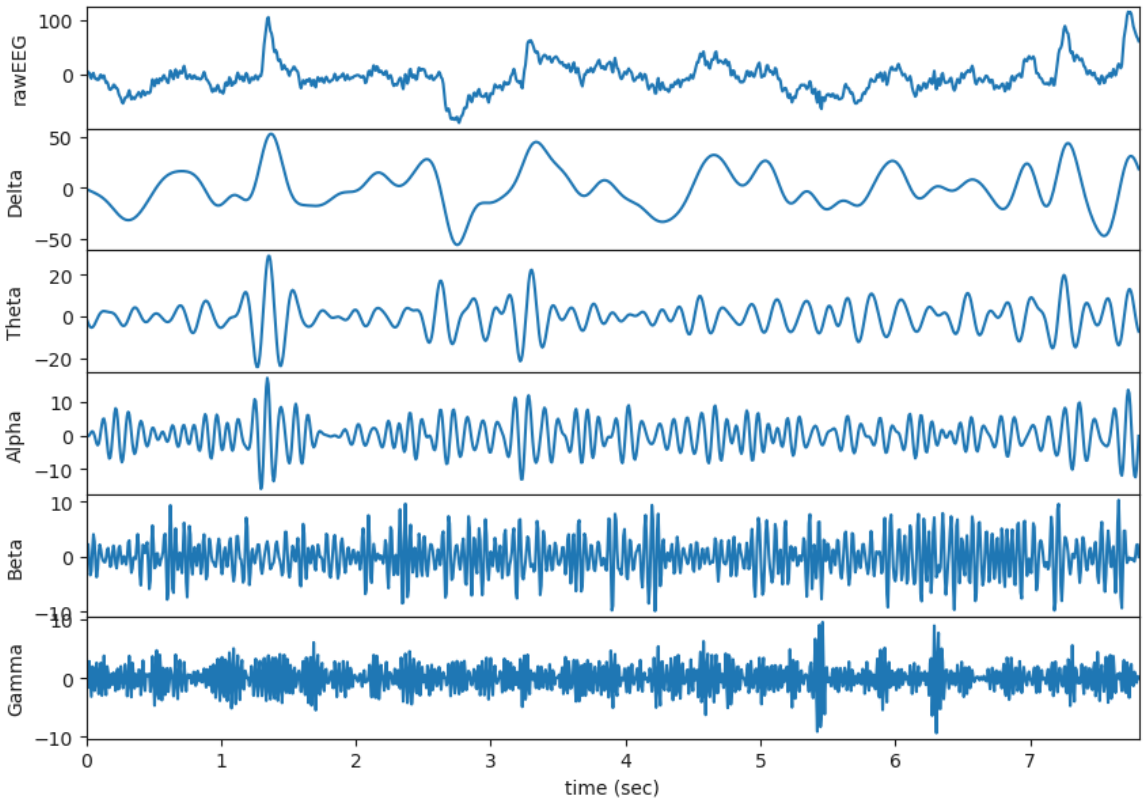


Fig. 1.1 The signal channel raw EEG signal and corresponding frequency bands: Delta (0.1 – 4 Hz), Theta (4 – 8 Hz), Alpha (8 – 14 Hz), Beta (14 – 30 Hz), Gamma (30 – 63 Hz)

Delta

Delta waves were first introduced by Walter in 1936, it ranges from 0.5 to 4 Hz in frequency. Delta waves are usually observed in deep sleep. Since delta wave is the low frequency wave, it is easily confused by the movement artifact, due to similar nature. Delta waves have also been linked to continues attention tasks.

Theta

Theta waves were introduced by Dovey and Wolter, ranges from 4 to 8 Hz in frequency. Theta waves are linked to drowsiness and deep meditation state.

Alpha

Alpha waves, perhaps are the most widely investigated waves in EEG studies. Alpha waves were introduced by Berger in 1929. They lie in a range from 8 to 14 Hz. Alpha waves usually

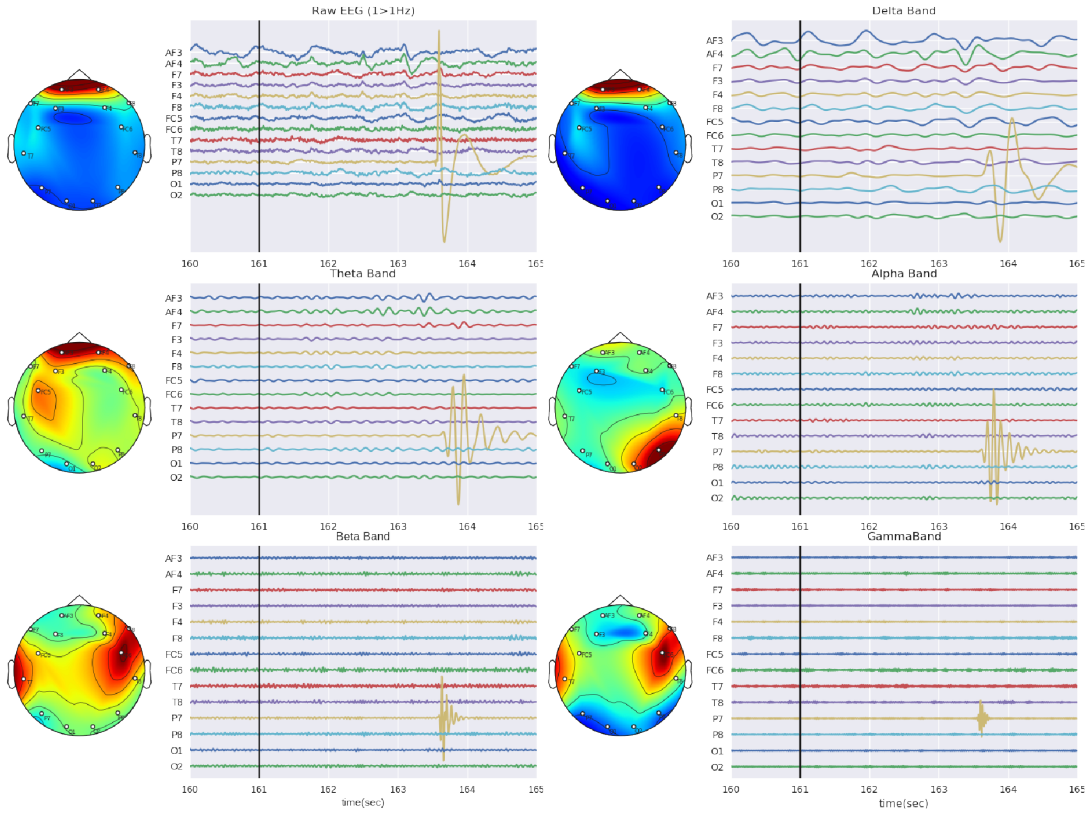


Fig. 1.2 A 14-channel EEG signal and corresponding brain activity in different frequency bands.

appear on the occipital lobe of the brain. Alpha waves are a most common indication of relaxing state of mind, and also linked to closing eyes. Any sign of anxiety or attention reduces the alpha waves.

Beta

Beta waves lie in the range of 14-30 Hz of frequency. Beta waves have been associated with active thinking, anxious, high alert, and focus of the brain.

Gamma

Gamma waves are the higher frequency waves, ranges from 30 to onwards. Gamma wave is considered as to play a complex role in brain functionality, such as combining information from two different sensory inputs. It is also used to confirm certain brain diseases.

The electrode placements on the scalp are done according to a 10-20 system. Each electrode is labeled according to the placement on the scalp. The labels start with a character; Frontal (F), Temporal (T), Parietal (P), Occipital (O) and Central (C).

1.5 Quantifying the auditory attention on natural speech

The quantification of cognitive attention is not an easy task. It is certainly very difficult for visual attention. However, for the auditory attention, the decades of research have evidently established the formulation for quantification. For experimental design based on an odd-ball paradigm, in which a series of sinusoidal tones are presented to the participant, and the participant is asked to respond on the odd-ball tune (tune of odd frequency). The attention is computed, how many times the participant correctly identified. Similar Odd-ball paradigm is also used for visual attention, in which participant is presented series of pictures from a similar category (car, train, bike) and ask to respond on an odd picture (face). In this case of visual attention, attention can be quantified as well.

For auditory attention natural speech, most widely accepted method to measure the attention level is to count the number of correctly identified words in shadowing task (dichotic listening task) [16–19]. For our experiment, we have used a similar approach to compute the auditory attention score on natural speech.

1.6 Problem formulation

Processing physiological signals have been an established research topic for a long time. Understanding the relationship between physiological signals and psychological or behavioral processes [20] opened the door to many applications [21]. Recent advancements in recording the EEG and other physiological signals have made it easy to employ the devices for day-to-day activities of life. The auditory attention plays a vital role in learning and performance of the task. Sparingly, very limited studies have focused on investigating the auditory attention on natural speech.

In this thesis, we aim to investigate and quantify the auditory attention from physiological responses such as EEG, GSR, and PPG. The primary objective of work is to predict the

auditory attention level of a person from physiological responses. First, we design an experiment, based on a dichotic listening task, with different auditory conditions. Then we collect and label the physiological responses properly. The statistical analysis could be used to identify the significance of auditory conditions for attention level. The signal analysis can reveal the correlation between physiological responses and the attention level. Since auditory conditions influence the attention level, the correlation of physiological responses and auditory conditions can also be analysed.

The collected database then, are used to formulate the predictive tasks. Since there is a lack if the database in the literature for auditory attention. We aim to release the collected dataset in the public domain. Finally, advanced machine learning and signal processing techniques are used to improve the predictive tasks

1.7 Related work

As with any other sensory modality, attention plays an important role when processing auditory information and therefore can have an impact on language processing and learning. For instance, it is well known that task-irrelevant auditory stimuli can undermine short-term memory and have a negative impact on language comprehension and learning [22, 23]. In addition, it has been suggested that the mechanism of auditory selective attention might depend on the perceptual load [24]. Understanding auditory attention in non-native audiences is therefore critical for designing effective learning environments.

A major focus of research in auditory attention has been on understanding selective attention, change deafness and spatial attention. Auditory change deafness refers to the mechanisms by which the brain misses information in auditory stimuli during short time intervals. Auditory change deafness is analogous to the phenomenon of a psychogenic blackout or transient loss of consciousness [25] in that there is a short interval in continuous listening processes, during which the brain does not process any auditory information. Previous auditory attention studies have focused on the effect of the length of stimulus and the complexity of sentence [16] and background noise [26]. The authors in [16] demonstrated that simple structural English sentences were more easily perceived than complex structural ones, however, the length of the sentence did not have a considerable impact on it. In [26] it was concluded that the effect of noise on a perception of speech diminishes with increased signal to noise ratio (SNR).

Among physiological signals, EEG, GSR or Electrodermal activity (EDA), Electromyography (EMG) and Electrocardiography (ECG) are widely used. Employing EEG signals to interface with computers is one of the most popular systems, known as Brain Computer

Interface (BCI). There is a considerable literature for BCI applications, ranges from clinical applications [27–30] to help patients to daily-life activities for playing games [31] and assessment of emotions [32–34]. Related studies are discussed in detail in Chapter 6. Other related studies are also discussed in respective chapters.

1.8 Organization of thesis

The thesis is organized as follow; In Chapter 2, we describe the methodology used to design and conduct the experiment. In the second half of the chapter, we introduced the formulation of predictive tasks. In the Chapter 3, we demonstrate the statistical analysis of the text responses collected from the experiment. Statistical analysis provides the insight into the effect of auditory conditions on attention score. In Chapter 4, we introduce an artifact removal algorithm based on wavelet packet analysis. The proposed algorithm has two tuning parameters and three operating modes. We demonstrate the performance of the proposed algorithm and compare it with the state-of-the-art algorithm. Chapter 5, provides the signal analysis of collected physiological responses. We show the correlation analysis, spectrum analysis and ERP analysis of EEG signals. Chapter 6 describes the collected database, released in the public domain. In Chapter 7, we exploit the spatial relationship of EEG electrodes using Convolutional Neural Network. The Chapter 8, we introduce a 3D difficulty model for serious game design. We conclude our work and indicate some future directions in Chapter 9.

Chapter 2

Methodology

In this chapter, we present the methodology of the experiment. First, we describe the experiment design including the description of stimuli and signals used. Secondly, we describe the procedure of experiment conduction and computation of attention score. Finally, we formulate four predictive tasks for the collected data.

2.1 Experiment design

The dichotic listening task is a popular approach in psychology that is used to investigate how the brain selects one stimulus over another. During the dichotic listening task, a subject is presented two independent speech stimuli in each ear simultaneously and has to attend to and write down one of the corresponding messages. Studies using the dichotic listening task have demonstrated that the brain is capable to focus on the target message, although the ability to remember its details is limited [6]. Our experiment design is inspired by the dichotic listening task and seeks to assess the level of auditory attention under different auditory conditions. However, unlike the dichotic listening task, our experiment design presents the same audio message in both ears, and different auditory conditions are created every time an audio message is presented. Based on previous studies on auditory information processing [16, 26], we create different auditory conditions by controlling the level of background noise, the semanticity of the audio message and finally its length. The participants are subjected to three consecutive tasks, namely listening, writing and resting. During the listening task, a participant is presented an audio message in a specific auditory condition. Participant is subsequently asked to transcribe the message during the writing task and a resting period ensues.

Three physiological responses namely EEG, GSR and PPG recorded at a sampling rate of 128 Hz for the entire experiment. The attention score was computed by counting the

number of correctly transcribed words for each trial for each participant. The schematic of the experiment design is shown in Figure 2.1. In Figure 2.1, physiological responses are labeled as *R*, attention score as *A* and auditory conditions as *Nl*, *Sm* and *Ls* for background noise level, semanticity, and length of stimulus respectively. The physiological responses were properly labeled with the corresponding task (listening, writing, resting), attention score and auditory condition.

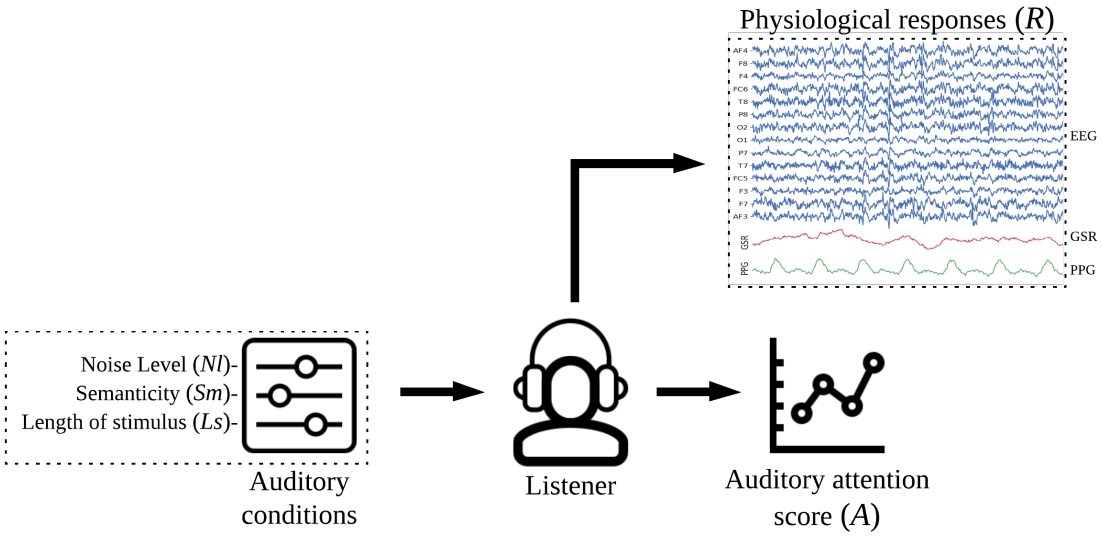


Fig. 2.1 Experimental model

2.2 Audio stimuli

2.2.1 Audio dataset

A total of 5000 audio clips were obtained from the Tatoeba Project [35] along with their corresponding text. The Tatoeba Project is an open online multilingual sentence dictionary. All collected audio clips were semantically correct English sentences of lengths ranging from 3 words to 13 words per sentence. To ensure the consistency of physical attributes (e.g. pitch, rate of speech etc) of stimuli, the selected audio stimuli were reproduced by the same voice.

Generation of non-semantic stimuli

A collection of 1700 non-semantic audio stimuli was generated from the original semantic stimuli. Non-semantic audio stimuli were generated by suitably inserting in one (or more)

audio clip(s) of isolated words reproduced by the same speaker. The timing of the insertion was determined by identifying long enough silence intervals in the original audio clip. The transition between original words and new inserted isolated words was ensured to be smooth. The resulting length of generated non-semantic stimuli also ranged from 3 to 13 words. The following sentences **Tx1**, **Tx2**, **Tx3** and **Tx4** are examples of semantic and non-semantic sentences:

Tx1: *I am going to study.*

Tx2: *I would like to read some books about the Beatles.*

Tx3: *Let's* ***touch enjoyable*** *go.*

Tx4: *I have a* ***hey*** *big* ***are we*** *dog.*

Among these examples, **Tx1** and **Tx2** are two original semantic stimuli collected from Tatoeba Project [35], whereas **Tx3** and **Tx4** are non-semantic stimuli generated from semantic ones. Specifically, sentence **Tx3** is made by inserting the highlighted words *touch* and *enjoyable* into an original semantic sentence; *Let's go*. This approach of generating non-semantic sentences is based on [36], and was adopted to mimic a real-world scenario, where a few unknown or unrelated words in a sentence make it appear non-semantic to the listener. This situation is commonly experienced by non-native speakers as a consequence of a more limited vocabulary.

Background noise in stimuli

Both groups of stimuli, semantic and non-semantic were used to create six sets of stimuli by adding different levels of background noise. The noise produced by talking crowds was used as background noise and ensured not to be identical for all the stimuli. The resulting six sets of stimuli have signal-to-noise ratio (SNR): -6 dB, -3 dB, 0 dB, 3 dB, 6 dB and ∞ dB (noise free). This procedure resulted in six sets of stimuli, each consists of two groups, 5000 semantic and 1700 non-semantic stimuli. Since the same audio stimuli are used to generate noisy stimuli, the audio messages in each set are identical, except the noise level.

Length categorization of stimuli

All the stimuli were grouped according to their length into three categories, namely small (L1), medium (L2) and long (L3). The average lengths of L1, L2, and L3 are 4, 8 and 12 words respectively with a variation of ± 1 word. For length categorization, we followed closely the approach presented in [16], which investigated aural perception. In allowing

different sentence lengths within each group, we recognized that some words are phonetically longer than others, i.e *'I'* and *'Congratulation'*, and assumed that a difference in sentence length of one word does not have any significant impact on the listening task [37] [38].

2.2.2 Audio selection

Our audio dataset consisted of stimuli presenting different noise levels, length and semanticity. While conducting the experiment, each participant was presented with 144 stimuli, 72 from the semantic group and 72 from the non-semantic one. Each group, semantic and non-semantic, consisted of 30, 24 and 18 stimuli from small (L1), medium (L2) and long (L3) categories. In addition these categories were equally divided into six sub-categories of noise levels i.e 30 / 6 = 5, 24 / 6 = 4 and 18 / 6 = 3 from each sub-group for L1, L2 and L3 respectively, as shown in Figure 2.2.

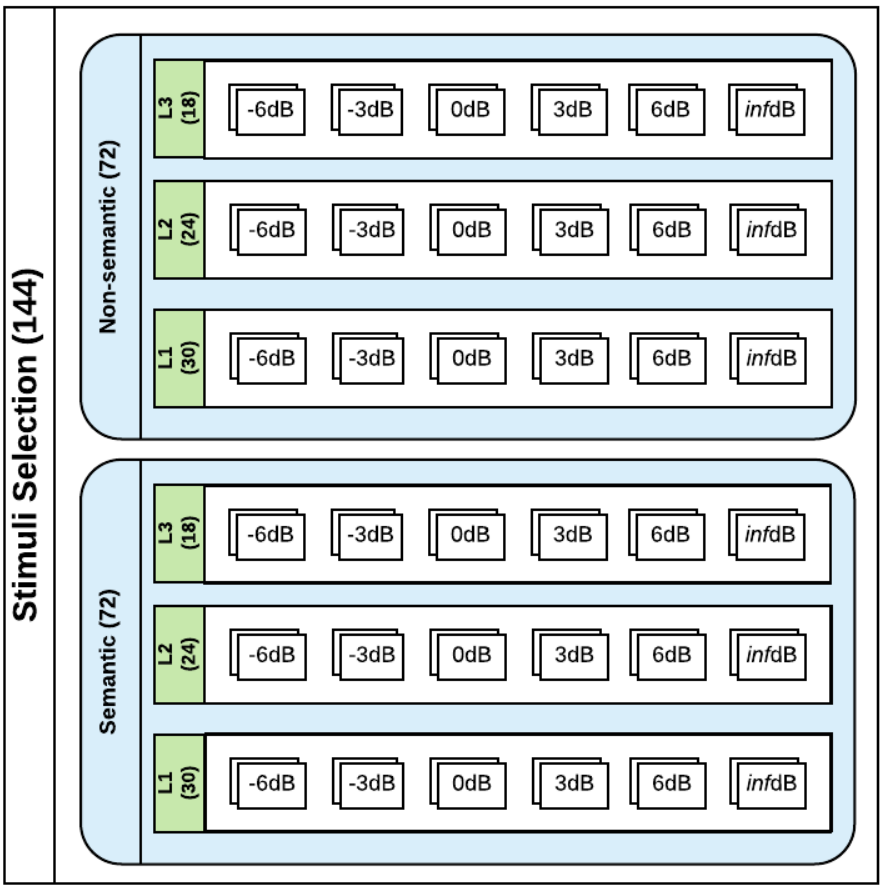


Fig. 2.2 Selection of Stimuli for each participant.

The selection of stimuli for a single participant was random without replacement, ensuring no repetition of the same audio message, even with a different level of noise. This combination of a different level of noise, semanticity, and length of stimuli resulted in 36 different experimental conditions ($6 \times 3 \times 2 = 36$). The number of stimuli per experimental condition only differed with the length of stimuli, which is 5 for small (L1), 4 for medium (L2) and 3 for long (L3), as shown in Table 2.1.

For readability purposes, each experimental condition is labeled as *'xxdByLz'* in this thesis, where *xxdB* indicates the noise level, *y* is a binary digit 0 or 1 indicating semanticity or non-semanticity respectively, and *Lz* corresponds to the length of the sentence, namely L1, L2 or L3. For example, the label *-3dB0L2* indicates the experiment condition with -3 dB SNR for semantic stimuli of length L2.

Table 2.1 Number of stimuli per experimental condition.

SNR	**Semantic**			**Non-Semantic**		
	L1	**L2**	**L3**	**L1**	**L2**	**L3**
-6 dB	5	4	3	5	4	3
-3 dB	5	4	3	5	4	3
0 dB	5	4	3	5	4	3
3 dB	5	4	3	5	4	3
6 dB	5	4	3	5	4	3
∞ dB	5	4	3	5	4	3
Subtotal	30	24	18	30	24	18
Total	72			72		

2.3 Physiological signals

For physiological responses, we recorded three different signals, namely EEG, GSR, and PPG. The methods and apparatus to record these signals are described subsequently.

Electroencephalogram signals

To record EEG signals, an Emotiv Epoc wireless headset device was used [39]. This headset provides 14 EEG channels with electrodes position as in a 10-20 system. The electrode positions are *AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1*, and *O2*. The device and electrode positions are shown in Figure 2.3. The maximum sampling rate supported by the device is 128 Hz. The signals were sampled at a rate of 128 Hz and recorded using API provided by the manufacturer of the device. While recording, raw EEG signals were filtered

with an Infinite Impulse Response (IIR) filter to remove the high DC component present. The implemented filter is described by the difference equation;

$$y_e(n) = \frac{\alpha - 1}{\alpha}(x_e(n) - x_e(n-1) + y_e(n-1)), \tag{2.1}$$

where $x_e(n)$ is a single channel raw EEG signal, $y_e(n)$ is corresponding filtered signal. We used $\alpha = 256$, which provides around 20 dB of attenuation for the DC component, and has negligible effect on the AC part [40]. The frequency response of the implemented filter is shown in Figure 2.4.

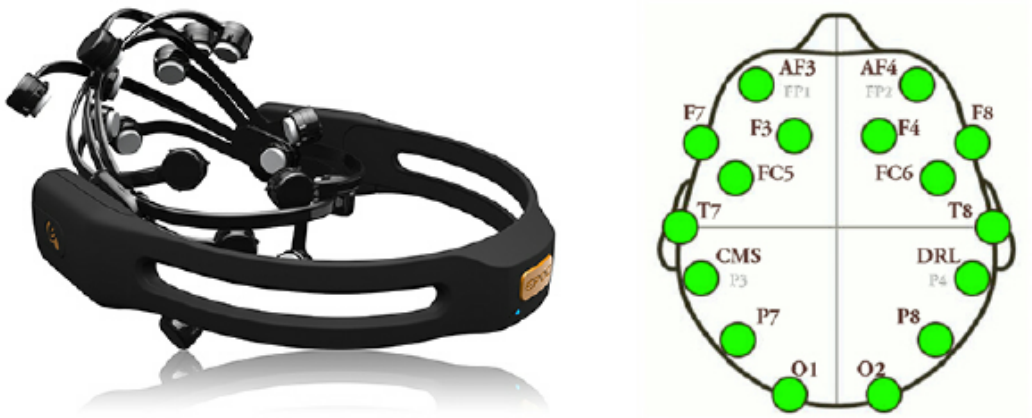


Fig. 2.3 Emotiv Epoc 14 channel and electrode position map (10-20 system), source of image: www.emotiv.com

Galvanic Skin Response signals

The state of the sweat glands in the skin modulates the conductance of the skin and is controlled by the sympathetic nervous system, which is measured electrically as GSR, also named Electrodermal Activity (EDA). Two small metal plates placed on the index and middle fingers were used to measure the skin conductance, as shown in Figure 2.5. The metal plates were connected through Arduino to interface with the computer. While recording the GSR signal, it was pre-processed at Arduino by a moving-average filter characterised by the equation;

$$y_g(n) = \frac{1}{N}\sum_{k=0}^{N-1} x_g(n-k) \tag{2.2}$$

where $x_g(n)$ is the raw GSR signal and $y_g(n)$ is the filtered GSR signal. The number of samples chosen to be averaged, was 100, to ensure minimal delay. The raw signal; $x_g(n)$ and

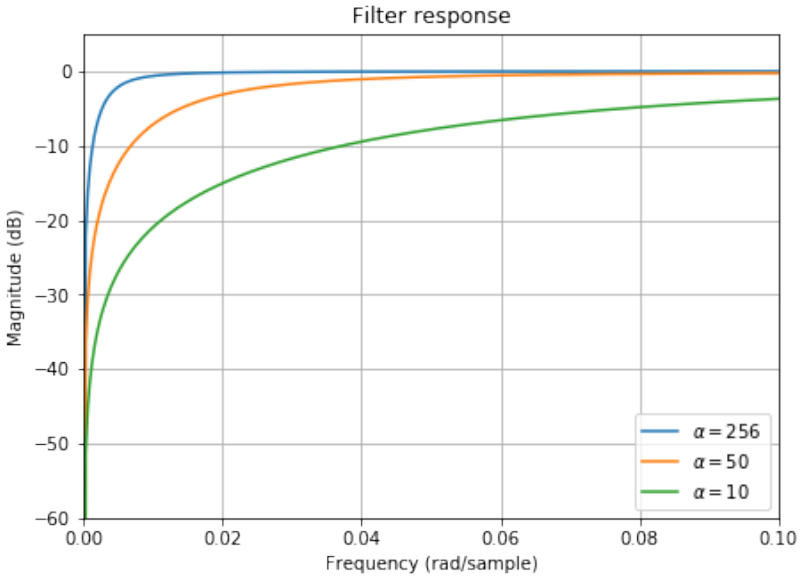


Fig. 2.4 Frequency response of IIR filter

filtered signal; $y_g(n)$, were recorded at a sampling rate of 128 Hz and stored along with EEG signals.

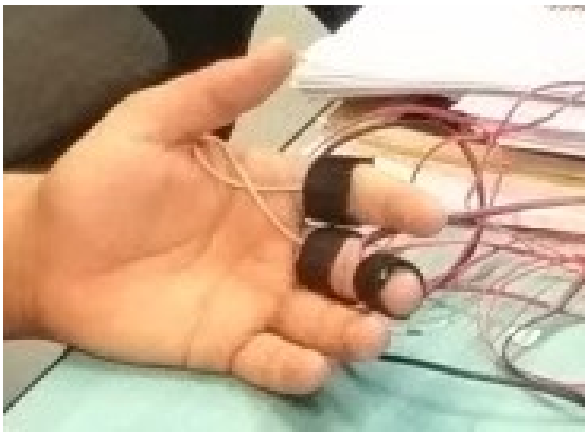


Fig. 2.5 Sensors (metal plates) for Galvanic Skin Response on index and middle finger.

Photoplethysmogram signal

For recording the photoplethysmogram response, a pulse sensor [41] was used as shown in Figure 2.6. Studies have shown the photoplethysmogram measure is close to electrocardiogram (ECG) response [42]. The pulse sensor was placed on the ring finger and interfaced through a separate Arduino to record the signal. From the raw signal, the pulse rate and the

interbeat interval (IBI) were estimated using source code provided by the sensor manufacturer [43]. These three signal streams were recorded at a sampling rate of 128 Hz.

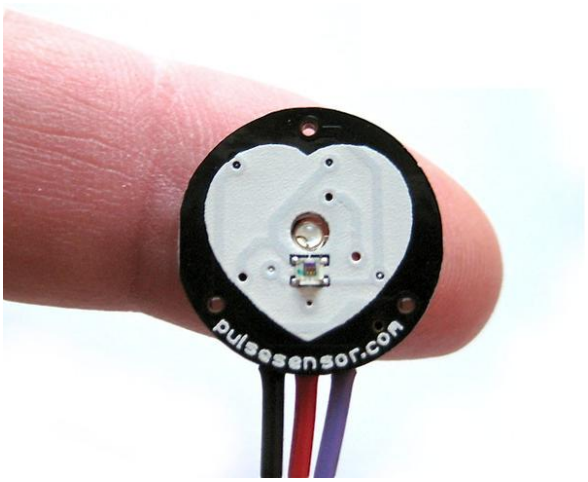


Fig. 2.6 Pulse sensor, source of image: www.pulsesensor.com

All the physiological signal streams (EEG, GSR, and PPG) were recorded at a sampling rate of 128 Hz. In total, 19 signals streams were recorded i.e. 14 EEG channels, 2 GSR streams (raw signal and filtered signal) and 3 PPG streams (raw signal, pulse rate, and IBI). While recording, all the signals were properly labeled to identify the task (listening, writing or resting) and the auditory condition (noise level, semanticity, and length of stimulus) presented to the participant. For each trial, the text transcribed by participants and the original text were recorded.

2.4 Experiment conduction

2.4.1 Participants

A group of 25 healthy, university students of science and technology, with no known auditory processing disorder participated in the study. There were 21 male and 4 female participants, and all the participants were non-native English speakers (i.e, their first language was other than English). We chose non-native speakers in order to measure attention in a non-trivial task, which is typical of learning. The participants came from different nationalities (see Table 2.2a) and different first language (see Table 2.2b). In some cases, participants sharing their nationality presented different first languages (e.g. Indian) and in other cases, participants from different nationalities presented the same first language (e.g. Arabic). The age distribution of the group of participants is shown in Table 2.3.

Table 2.2 Distribution of nationality and first language of participants

(a) Nationality

Nationality	**Number of participants**
Algerian	1
Indian	8
Iranian	3
Italian	4
Kazakh	1
Lebanese	4
Moroccan	1
Nepalese	1
Pakistani	1
Tunisian	1

(b) First language

First language	**Number of participants**
Arabic	7
Farsi	3
Italian	4
Kannada	1
Kazakh	1
Mathili	1
Malayalam	4
Marathi	1
Tamil	1
Telgu	1
Urdu	1

Table 2.3 Distribution of age group

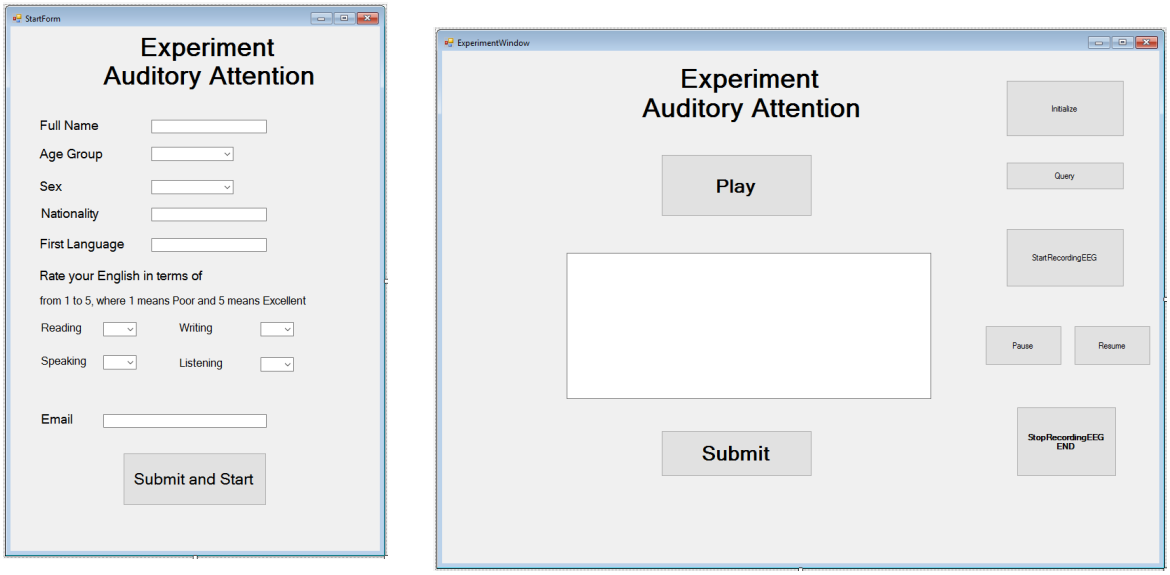
Age group (years)	**Number of participants**
16-20	1
21-25	6
26-30	16
31-35	2

2.4.2 Procedure

Participants were presented with a computer interface, designed in C# as shown in Figure 2.7. After mounting all the sensors (EEG, GSR, and PPG) on a participant, a passive earphone was provided for listening. With the computer interface window 1 (Figure 2.7a), participants were asked to enter basic demographic information, namely sex, nationality, age-group, and first language. Participants were also asked to rate their English language skills in terms of Reading, Writing, Listening and Speaking on the scale of 1 to 5, with 1 being poor and 5 being excellent.

After submitting their demographic information (Figure 2.7a) a window opened for the actual experimental tasks (Figure 2.7b). The participants could initiate the listening task by clicking the play button. During the listening task, the computer interface remained disabled so as to prevent participants from engaging in other activities. Once the audio file had finished no more reproductions were permitted and participants were allowed to submit a

transcription of the audio sentence. Upon submission, participants could reproduce the next audio file by clicking the play button again. The pictorial representation of the procedure on timeline is shown in Figure 2.8. The duration between submission of transcription and playing the next audio stimulus is labeled as resting. Each trial starts with a click of the play button and ends with the next click of the play button. A participant could not replay the same audio file. This procedure was carefully explained to each participant beforehand. A picture of a participant demonstrating the conduction of the experiment is shown in Figure 2.9. On average, the total procedure involving the 144 audio stimuli (144 trials) described in Section 2.2.2 took 35 minutes. All the collected responses are anonymised properly for any further analysis and use.



(a) Interface for entering demographic information

(b) Interface for experimental tasks

Fig. 2.7 Computer interface for experiment

2.5 Attention score

The attention score of participants for each stimulus was computed following previous studies [16–19] i.e. by counting the number of correctly identified words in the transcribed text. The attention score for the p^{th} participant in the k^{th} experimental condition for the i^{th} stimulus is computed as follows:

$$A_{k,i,p} = \frac{N_{C(k,i,p)}}{N_{T(k,i,p)}} \times 100 \tag{2.3}$$

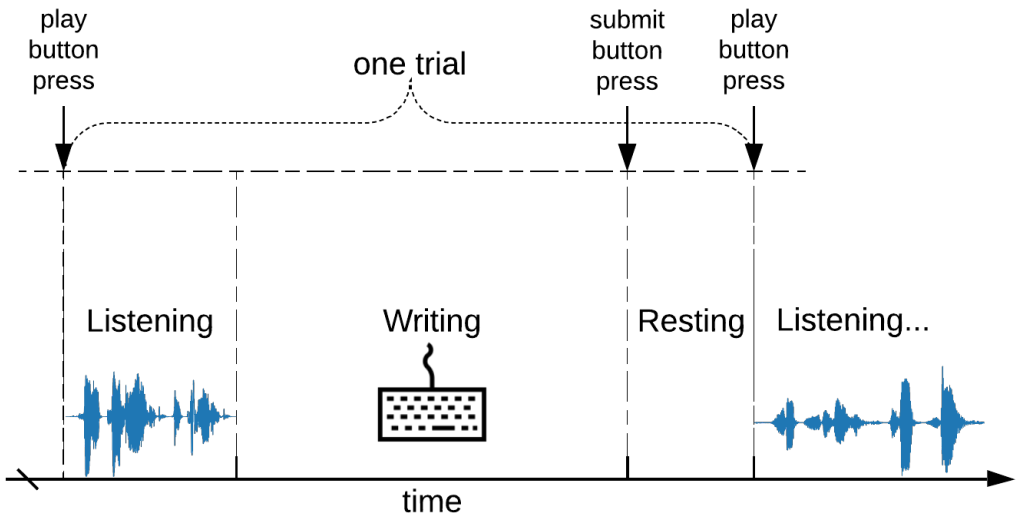


Fig. 2.8 Timeline of experimental procedure, showing listening, writing and resting segments with alone with events of button pressing.

Fig. 2.9 A participant, while experiment procedure. A consent of participant was taken to use the picture to display the experiment procedure.

where $N_{C(k,i,p)}$ is the number of correctly identified words in the i-th text response and $N_{T(k,i,p)}$ is the total number of words in the i-th original text of stimulus for the k-th experimental condition and the p-th participant. While computing the score, minor errors in spelling and typos were ignored. For example *beautiful/beutiful*, *does/dos*, *dogs/dog*.

2.6 Analysis approaches

The collected responses, labeled against experiment condition and attention score open the door to various types of analysis. We will focus our analysis in a few directions that are explained here. First, we consider the statistical analysis of the attention score and auditory

conditions, which is explained in Chapter 3. Next, we analyse the recorded signals and preprocessing algorithms which are explained in Chapter 4 and 5. Finally, we exploit the findings of text response analysis and preprocessing and carry out the predictive analysis of attention score and auditory conditions from physiological responses, which is explained in Chapter 7.

2.7 Formulation of predictive tasks

Based on the experimental model shown in Figure 2.1, we formulate four predictive tasks as described in the following subsections. For all the predictive tasks, the input is the physiological responses R. First, we describe two ways of feature extraction from physiological responses. A segment of the signals Sg is defined as a segment of physiological responses R for the entire duration of a particular sub-task, e.g. listening, writing and resting denoted as Sg_l, Sg_w and Sg_r respectively. All the segments can be obtained from physiological responses R ($R \rightarrow Sg_l, Sg_w, Sg_r$), then the features F_r can be extracted in two ways, namely segment wise and window wise. In segment wise feature extraction, the entire segment of a sub-task is used (e.g $Sg \rightarrow F_r$), as shown in Figure 2.10. In Figure 2.10, a segment Sg is highlighted with a box. For window wise feature extraction, smaller overlapping windows, (size of a window, $win <$ size of a segment, Sg) are used to extract the features ($Sg[win] \rightarrow F_r$), as shown in Figure 2.11. In Figure 2.11, a selected window is shown in a box, whereas the rest of the segment Sg is shaded with blue.

2.7.1 Task 1: Attention score prediction

The objective of this task is to predict the auditory attention A score of a listener from the physiological responses R, and can be formulated as

$$A \approx f(R).$$

Since the attention score is computed for the auditory process and collected physiological responses are labeled with corresponding attention score for each listening segment Sg_l, the features F_r can be extracted from each listening segments ($R \rightarrow Sg_l \rightarrow F_r$). A function f can be modeled such that the predicted attention score A' for test data (unseen data) can be estimated as

$$A' = f(F_r)$$

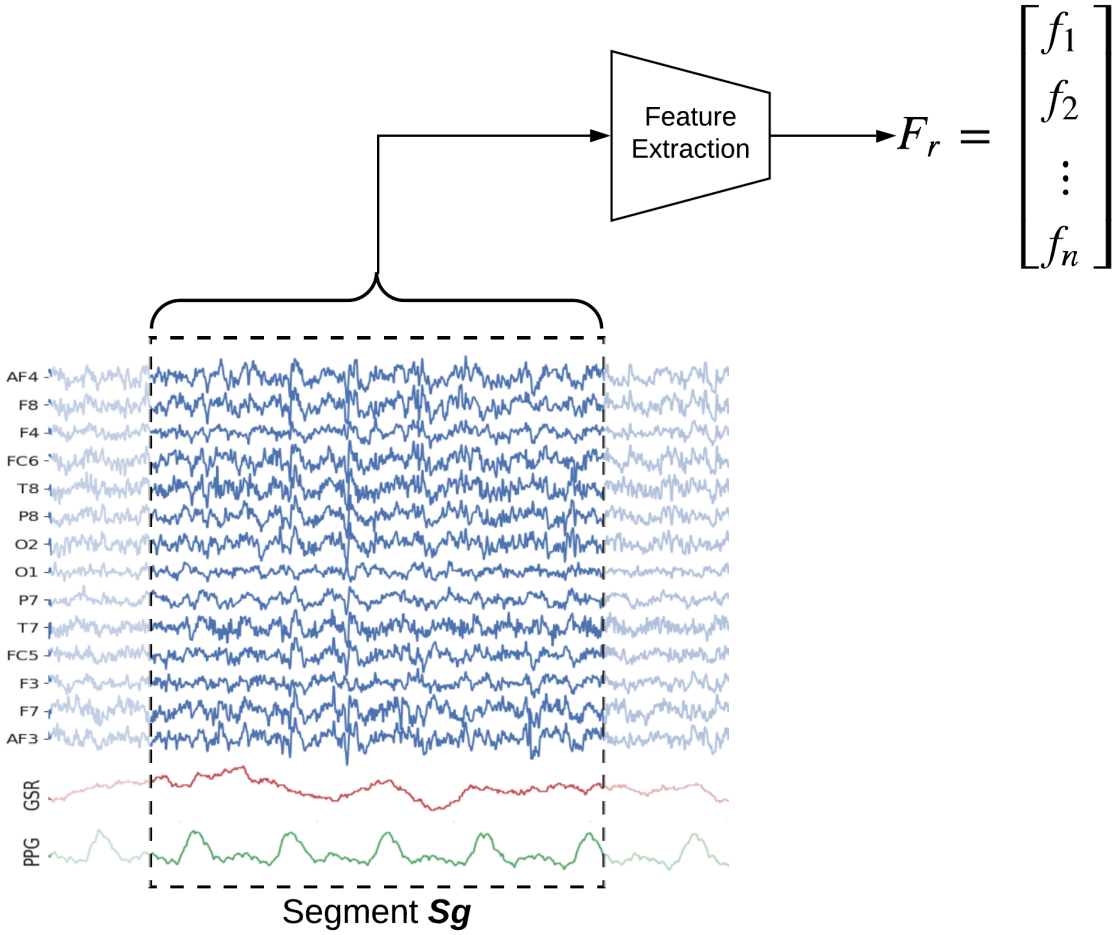


Fig. 2.10 Feature extraction from a segment Sg

by solving

$$\min_{f} \mathcal{E}(f)$$

where $\mathcal{E}(f)$ is expected risk

$$\mathcal{E}(f) = \mathbb{E}[\mathcal{L}(A, f(F_r))]$$

and $\mathcal{L}(\cdot,\cdot)$ is loss function and $\mathbb{E}[\cdot]$ is expectation operator. As the attention score is a real valued number, ranges from 0 to 100, the choice of loss function can be Mean Square Error (MSE), Mean Absolute Error (MAE) or combination of two as Huber loss.

Since an attention score associated with a listening segment depends on the entire segment, only segment wise feature extraction is applicable for this task. However, considering the temporal behavior of the attention, accumulated features extracted from smaller overlapping windows (as shown in the Figure 2.11) can be used for sophisticated temporal models like

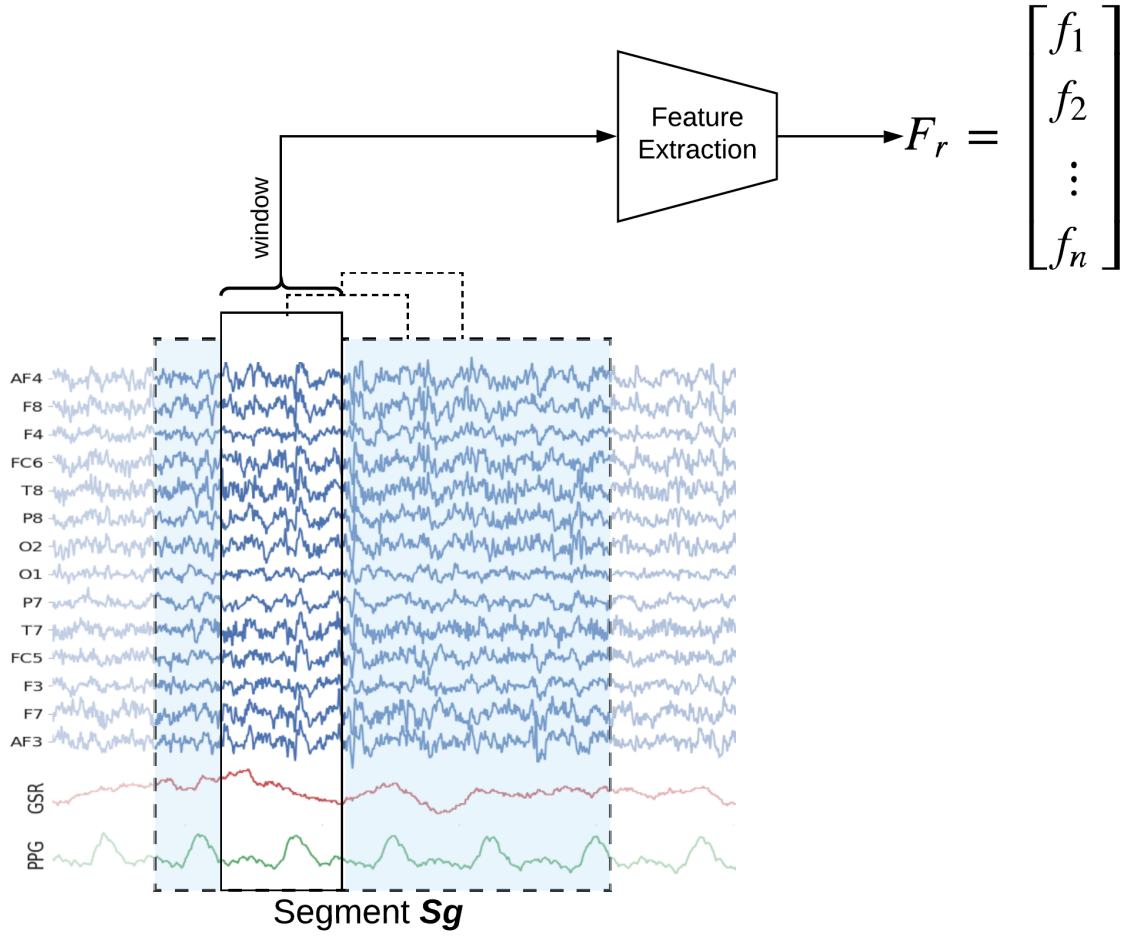


Fig. 2.11 Feature extraction from a small windows of a segment *Sg*

Bayesian Network and Recurrent Neural Network (RNN) with Long-Short Term Memory (LSTM) units.

Considering the statistical dependency (explained in Chapter 3) of attention score A on auditory conditions, as $A \approx f(Nl, Sm, Ls)$, the auditory conditions experienced by a listener can be predicted from the physiological responses, which can further be used to predict the attention score A with a hierarchical modeling approach. Noise level and semanticity of stimulus prediction are discussed in the next subsequent subsections. Predicting the length of stimulus is a trivial job, as the exact length of stimulus can be estimated from the length of recorded signal with known sampling rate.

2.7.2 Task 2: Noise level prediction

Considering the noise level (Nl) experienced by a listener stimulates the underlying effect in physiological responses (R), the noise level can be estimated as $Nl = f(R)$. As participants were listening to the audio with earphones in a controlled setting, it is fair to assume that the noise level experienced by listeners is the same as presented in the audio stimuli. To predict the noise level experienced by a listener, features F_r can be extracted from each listening segment ($Sg_l \rightarrow F_r$) and model a function f with corresponding noise level Nl such that predicted noise level Nl' for test data can be estimated as

$$Nl' = f(F_r)$$

by minimizing the expected risk $\mathcal{E}(f)$ over function f with appropriate loss function $\mathcal{L}(\cdot,\cdot)$. There are six levels of noise, used for the experiment, namely; -6 dB, -3 dB, 0 dB, 3 dB, 6 dB and ∞ dB (noise free). This predictive task can be considered as a regression by converting ∞ dB to some value greater than 6 or classification with 6 different classes. In the case of regression, choice of loss function can be similar to attention score prediction e.g. MSE, MAE or combination. However for classification, loss choice can be cross entropy loss.

Since the noise is presented in entire audio stimulus with the same level, features can be extracted from the entire listening segment ($Sg_l \rightarrow F_r$) or with smaller overlapping windows ($Sg_l[win] \rightarrow F_r$) as explained above. The size of the window can be considered as 1 sec (128 samples) with 50% overlapping, which allows predicting the noise level for a given short duration of physiological responses.

2.7.3 Task 3: Semanticity prediction

Similar to the noise level, semanticity Sm of stimulus experience by a listener can be considered to stimulate the underlying effects in physiological responses R; thus semanticity can be estimated as $Sm = f(R)$. Unlike the noise level, semanticity of stimulus is a subjective attribute. The statistical results (discussed in Chapter 3) shows the significant ($pvalue << 0.001$) difference in the performance of semantic and non-semantic stimuli, so it is assumed that participants experienced the same level of semanticity in the presented stimuli as labeled.

Similar to other tasks, for predictive modeling, the features F_r can be extracted from each listening segment ($Sg_l \rightarrow F_r$) and used to model a function f with corresponding semanticity label Sm such that the predicted semanticity Sm' for test data can be estimated as

$$Sm' = f(F_r)$$

by minimizing the expected risk $\mathcal{E}(f)$ of function f with appropriate loss function $\mathcal{L}(\cdot,\cdot)$. As each stimulus is labeled with a binary digit (0-semantic and 1-non-semantic), the task is binary classifications, and choice of loss function can be hinge loss or binary cross entropy loss. Since semanticity is the construct of complete stimulus, feature extraction from smaller windows is not applicable for this task, however similar to task 1, attention score prediction, a temporal model can be used.

2.7.4 Task 4: LWR classification, subtask prediction

Since the physiological responses *R* are labeled with listener subtasks (listening, writing and resting), a model can be designed to predict the subtask, that listener is performing. The features can be extracted for all the segments ($Sg_{(l/w/r)} \rightarrow F_r$) and can be used to model a function f such that the state of the listener can be predicted as

$$T' = f(F_r)$$

by minimizing the expected risk $\mathcal{E}(f)$ for unseen data, with appropriate loss function $\mathcal{L}(\cdot,\cdot)$. As a state of the task (e.g. listening, writing, resting) is labeled for the entire duration of the task, features can be extracted from an entire segment or from smaller windows, similar to task 2, which will allow predicting the state of a listener for a shorter duration of physiological responses. As there are three labels, this task is the classification task. We indicate this predictive task as *LWR*-classification task.

The summary of the four predictive tasks is tabulated in Table 2.4 below.

Table 2.4 Summary of predictive tasks

Prediction task	**Model**	**Feature extraction** $Sg \rightarrow F_r$	**Choice of loss function** $\mathcal{L}(\cdot,\cdot)$
Attention score	$A = f^*(F_r)$	Sg_l	MSE, MAE, Huber, Cross entropy, Hinge, Logistic
Noise level	$Nl = f(F_r)$		
Semanticity	$Sm = f^*(F_r)$		
LWR	$T = f(F_r)$	$Sg_{(l/w/r)}$	

* A temporal model like Bayesian network or RNN can be used with features from smaller windows.

Chapter 3

Text response analysis

In this chapter, we present the analysis of the text responses produced by participants during the experiment described in Chapter 2. First, we introduce the statistical methods used for analysis. Then, we discuss the results of the text response analysis.

3.1 Statistical methods

In this section, we present the statistical methods we have used to analyse the text responses collected during the experiments described in Chapter 2. In the experiment described in Chapter 2, a total of 25 participants were involved and were presented with 144 stimuli, characterised by six levels of noise, two degrees of semanticity and three different lengths, which resulted in 36 different experimental conditions. Each experimental condition was labeled as *'xxdByLz'* (e.g.-*3dB0L2* for -3 dB SNR, semantic stimuli of length L2).

3.1.1 Attention score computation

Attention level quantification remains a challenging task in psychology. The widely accepted method is to count the number of words correctly identified during a listening task [16–19]. Based on this method for quantifying the attention level, we define an attention score for each listening task that participants complete. Let $T_{i,p,k}$ denote the i^{th} transcription produced by the p^{th} participant under the k^{th} experimental condition. Then, its attention score, which we denote by $A_{i,p,k}$, was calculated as previously defined equation (2.3).

$$A_{i,p,k} = \frac{N_{C(i,p,k)}}{N_{T(i,p,k)}} \times 100,$$

where $N_{C(i,p,k)}$ is number of correct words in $T_{i,p,k}$ and $N_{T(i,p,k)}$ is number of total words in the original sentence. While counting correct words in transcription, minor errors in spelling and other typos were ignored, for example *looks/look*, *beautiful/beutiful* or *designed/disegned*.

For further analysing text response with respect to auditory conditions (e.g. experimental conditions), the average attention score $A_{p,k}$ of the transcriptions produced by the p^{th} participant under the k^{th} experimental condition was calculated as,

$$A_{p,k} = \frac{1}{I_k} \sum_{i=1}^{I_k} A_{i,p,k}, \tag{3.1}$$

where I_k is the total number of stimuli in k^{th} experimental condition (see Table 2.1). Each computed average attention score $A_{p,k}$ was used as an individual sample for statistical analysis. As $N_P = 25$ participants were involved in the experiment and each participant produced a total of $N_K = 36$ attention score samples (one sample per experimental condition), a total of 900 samples ($36 \times 25 = 900$) were available for further analysis.

3.1.2 Descriptive statistical analysis

The mean A_k and standard deviation S_k of the attention score for k^{th} experiment condition, across all the participants

$$A_k = \frac{1}{N_p} \sum_{p=1}^{N_p} A_{p,k} \tag{3.2}$$

$$S_k = \left(\frac{1}{N_p - 1} \sum_{p=1}^{N_p} (A_{p,k} - A_k)^2 \right)^{\frac{1}{2}}. \tag{3.3}$$

In addition, the mean attention score A_p for the p^{th} participant across all the conditions is computed as

$$A_p = \frac{1}{N_k} \sum_{k=1}^{N_k} A_{p,k}. \tag{3.4}$$

Finally, box-and-whisker, surface, and interaction plots were obtained for analyzing the impact of the noise level, the length of stimuli, and sentence semanticity on the attention score of transcriptions from each experimental condition.

3.1.3 Significance analysis

In order to analyse whether the independent variables noise level, semanticity, and length of stimulus have significant effects on the attention score, a repeated measure ANOVA test

[44] was applied. Then the student t-test was used to determine whether the mean of the attention score was significantly different under any two experimental conditions. Since a total of 36 experimental conditions are defined, the resulting p-values were represented in a 36×36 matrix, denoted as P-matrix. By definition, this P-matrix is symmetric and its diagonal represents the comparison of an experimental group with itself, hence the values in the diagonal are set to 1.0. In order to facilitate the analysis of each pair, heat maps were used to visually represent the p-values in the P-matrix, whose entries are labeled following the *xxdByLz* notation previously described. Threshold values of 0.05 and 0.001 were applied to the P-matrix, producing binary matrices in which significant differences with a 95% and 99% confidence level can be readily identified. The binary P-matrix allowed us to further arrange the experimental conditions in a hierarchical manner, following a bottom-up agglomerative method [45], where related experimental conditions are located close to one another. Further, the P-matrix was reordered by ranking the experimental conditions according to their mean attention score A_k.

3.1.4 Analysis of individual differences

In addition to comparing the attention score under different experimental conditions, in our study we analysed the individual variability under each experimental condition. Participants and experimental conditions were ranked according to their associated average correctness quantities,respectively A_p, as defined in (3.4), and A_k, as defined in (3.2). We used this ranking to produce a heatmap for the average correctness $A_{p,k}$. Further, for each participant, the mean and standard error of the attention score for each independent variable were computed.

3.2 Results and discussions

3.2.1 Impact of auditory factors

The mean A_k and standard deviation S_k of the attention score $A_{p,k}$ under each experimental condition are presented in Table 3.1. As expected, results show the attention score was lowest for low SNR and highest for high SNR. In general, an increase in the noise level produced a decrease in the attention score of transcribed sentences. In addition, irrespective of the noise level, an increase in the length of a sentence resulted in a decrease in the attention score as quantified from transcribed sentences.

The effects of noise level, length, and semanticity of stimulus on the attention score are shown in the box-and-whisker plots in Figure 3.1. The box-and-whisker plots indicate a

Table 3.1 Mean and standard deviation of the average correctness $A_{p,k}$ in each experimental condition.

	SNR	**Semantic**			**Non-Semantic**		
		L1	L2	L3	L1	L2	L3
Mean $(n = 25)$	-6 dB	13.03	5.86	4.85	10.41	8.21	6.91
	-3 dB	33.41	21.12	15.56	21.57	15.13	11.61
	0 dB	40.49	37.11	24.63	30.48	28.81	16.32
	3 dB	57.09	49.17	43.24	38.91	32.90	22.38
	6 dB	72.04	62.82	48.82	50.22	40.15	26.75
	∞ dB	85.17	86.48	72.80	67.03	56.20	39.45
SD $(n = 25)$	-6 dB	14.03	6.53	6.74	9.74	6.51	7.10
	-3 dB	23.02	17.02	8.21	14.96	10.04	7.40
	0 dB	20.27	22.63	20.09	15.00	18.06	12.10
	3 dB	24.88	26.21	25.74	18.17	19.19	14.54
	6 dB	21.34	23.28	24.80	20.25	18.36	15.86
	∞ dB	16.10	17.25	22.48	17.78	18.82	18.77

clear dependence of the attention score on the background noise level and the length of the stimulus. In addition, Figure 3.1 also suggests that the attention score is higher for semantic sentences than for non-semantic ones.

The effects of noise level and length of stimulus are also analysed separately for the semantic and non-semantic groups of stimuli. Interestingly, Figure 3.2 shows that the rate of change of the attention score is higher for semantic stimuli than non-semantic one. Specifically, the median of the attention score for the noise-free case (SNR = ∞ dB) was $median(A_{p,k\in k_1}) = 88$ in the semantic group, whereas for the non-semantic group is $median(A_{p,k\in k_2}) = 56$, where k_1 and k_2 are the sets of experimental conditions of noise-free case for semantic and non-semantic respectively. By contrast, in the noisy environment with the lowest SNR (-6 dB), the attention score was similar for the semantic and non-semantic groups. The dependence of the attention score on the length of a stimulus for the semantic and non-semantic groups is shown in Figure 3.3. Although a similar trend can be observed for both groups, the interquartile ranges and the medians were found to be different. Specifically for length L1, the medians of the attention scores are $median(A_{p,k\in k_3}) = 49.5$ and $median(A_{p,k\in k_4}) = 34.33$ for semantic and non-semantic groups respectively, and for L3, the medians are $median(A_{p,k\in k_5}) = 26.77$ and $median(A_{p,k\in k_6}) = 15.22$ respectively. The sets k_3 and k_4 are the experimental conditions of length L1 for semantic and non-semantic group respectively. Similarly, the sets k_5 and k_6 are for length L3.

The surface plot of the attention score for semantic and non-semantic stimuli with the noise level and the length of sentence (number of words) are shown in Figure 3.4, which

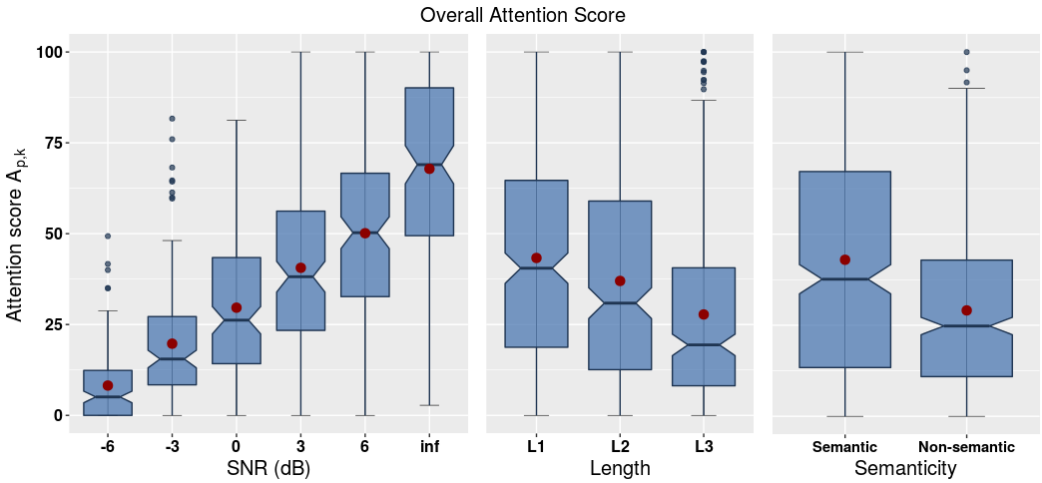


Fig. 3.1 Average attention score $A_{p,k}$ versus SNR, length and semanticity of stimulus.

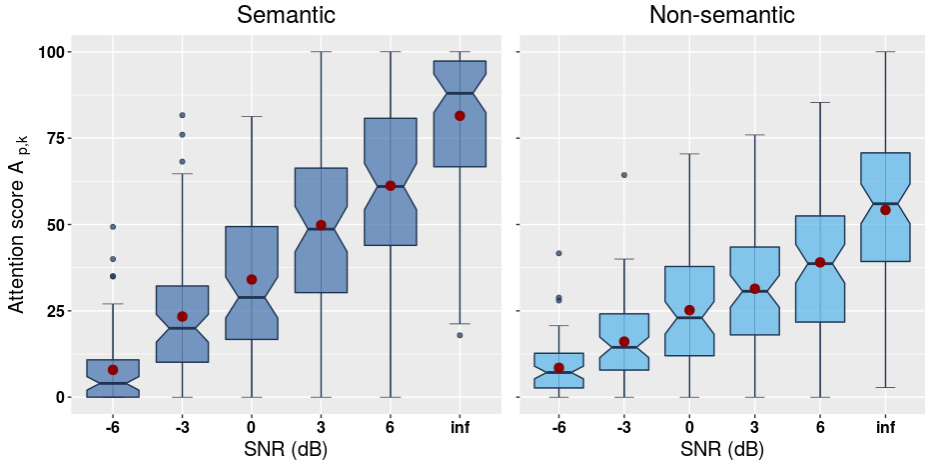


Fig. 3.2 Average attention score $A_{p,k}$ versus SNR for semantic and non-semantic sentences.

reveals that both planes are inclined towards a higher SNR and a smaller length of stimulus, from a lower SNR and a longer length of stimulus. As expected from other results, the plane corresponding to the non-semantic group is more inclined than the semantic one. It is interesting to notice, that the surface plan for the semantic group is overall slightly more elevated than the non-semantic group for a longer length of stimulus at higher SNR.

Figure 3.5a shows the impact of the noise level on attention score for semantic and non-semantic sentences. Interestingly, the effect of semanticity vanishes in high noise, whereas it is considerably high for high SNR. This effect can be explained by the Gap-Filling theory [46], a theory in cognitive neuroscience and linguistics according to which the brain is capable of predicting linguistic gaps by using syntactic or semantic priors [47] [48]. In other words, the brain can successfully use priors to fill in the gaps, whenever the sentences,

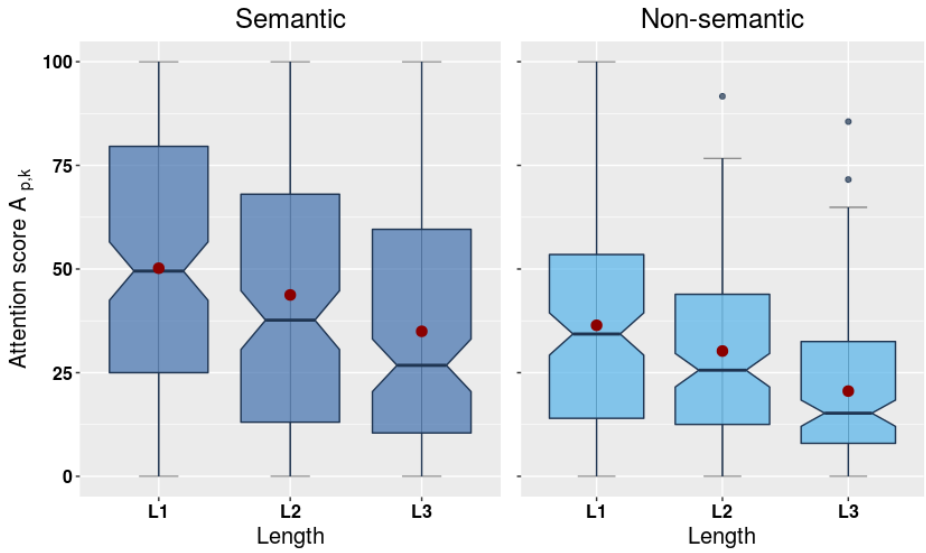
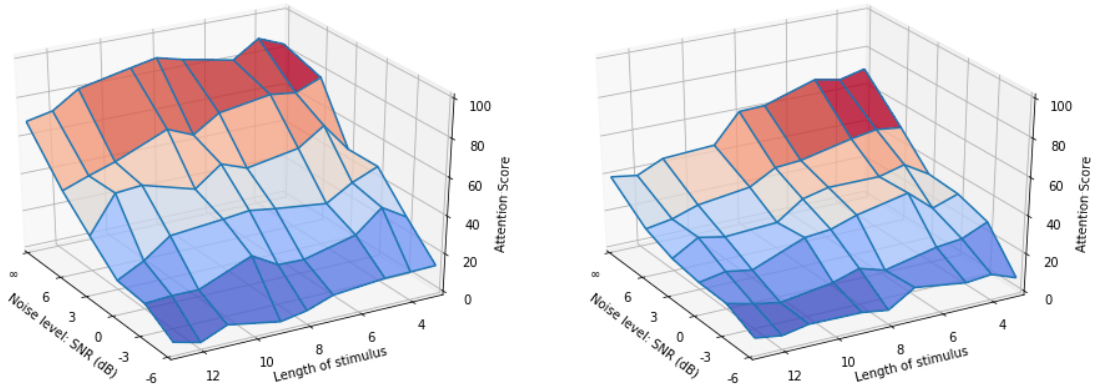


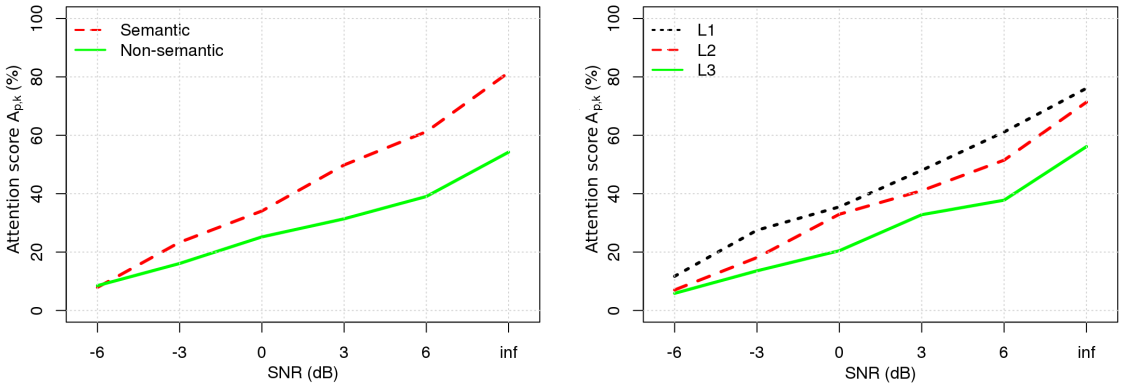
Fig. 3.3 Average attention score $A_{p,k}$ versus length for semantic and non-semantic sentences.



(a) Semantic (b) Non-semantic

Fig. 3.4 Attention score for semantic and non-semantic stimuli

irrespective to semanticity, are not audible clearly, as in noisy scenarios. That explains why for low SNR, the semantic sentences are as poor as non-semantic. In contrast, for a noise-free environment, the brain cannot use any priors since sentences are audible clearly, and the semantic sentences achieve a better attention score. Figure 3.5b shows the impact of noise on the attention score for different lengths of stimulus. As expected, shorter sentences have a more positive impact on attention than the medium length sentences followed by long sentences. Interestingly, our results suggest that this rate is similar for small and medium length sentences, and lower for large sentences.



(a) Semanticity and noise

(b) Length and noise

Fig. 3.5 Interaction between semanticity and length with noise.

Table 3.2 Results of the repeated measure ANOVA, where df denotes the degrees of freedom, SSq is sum of the squared differences, MSq is the mean sum of squares.

Source	df	SSq	MSq	F-value	P-value
Between	35	456255.11	13035.86	72.77	10^{-16}
Subject	24	121625.84	5067.74	28.29	
Within	864	272098.89	314.93		
Error	840	150473.04	179.14		
Total	889	728354.00			

3.2.2 Impact of experimental groups

The results from our descriptive analysis suggest that the attention score is affected by noise level, semanticity, and length of sentence.The results of repeated-measure ANOVA test are tabulated in Table 3.2, and indicate a strong evidence of a significant ($p \ll 0.001$ with $F(35,840) = 72.771$) difference between at least two experimental groups. The resulting partial eta-squared was $\eta^2_{partial} = 0.752$, which indicates that 75.2% differences in the experimental groups were due to the different experimental conditions. The ANOVA test was also performed on the noise level, semanticity and length of stimulus individually. The results for noise level ($p \approx 10^{-16}$, $F(5,120) = 238.500$) and for semanticity ($p \approx 10^{-10}$, $F(1,24) = 115.611$) and for length ($p \approx 10^{-16}$, $F(2,48) = 95.9220$) indicate a significant effect of these three auditory conditions on attention score individually.

Figure 3.6 shows the 36×36 P-matrix for the pairwise t-test analysis, as explained in Section 3.1.3. The experimental conditions which are represented as entries of the P-matrix, are arranged in order of first noise levels, followed by semanticity and then length. As expected by its definition, the P-matrix is symmetric and the values of the diagonal are 1. It is apparent that a major fraction of pairwise comparisons has a low p-value, which suggests that

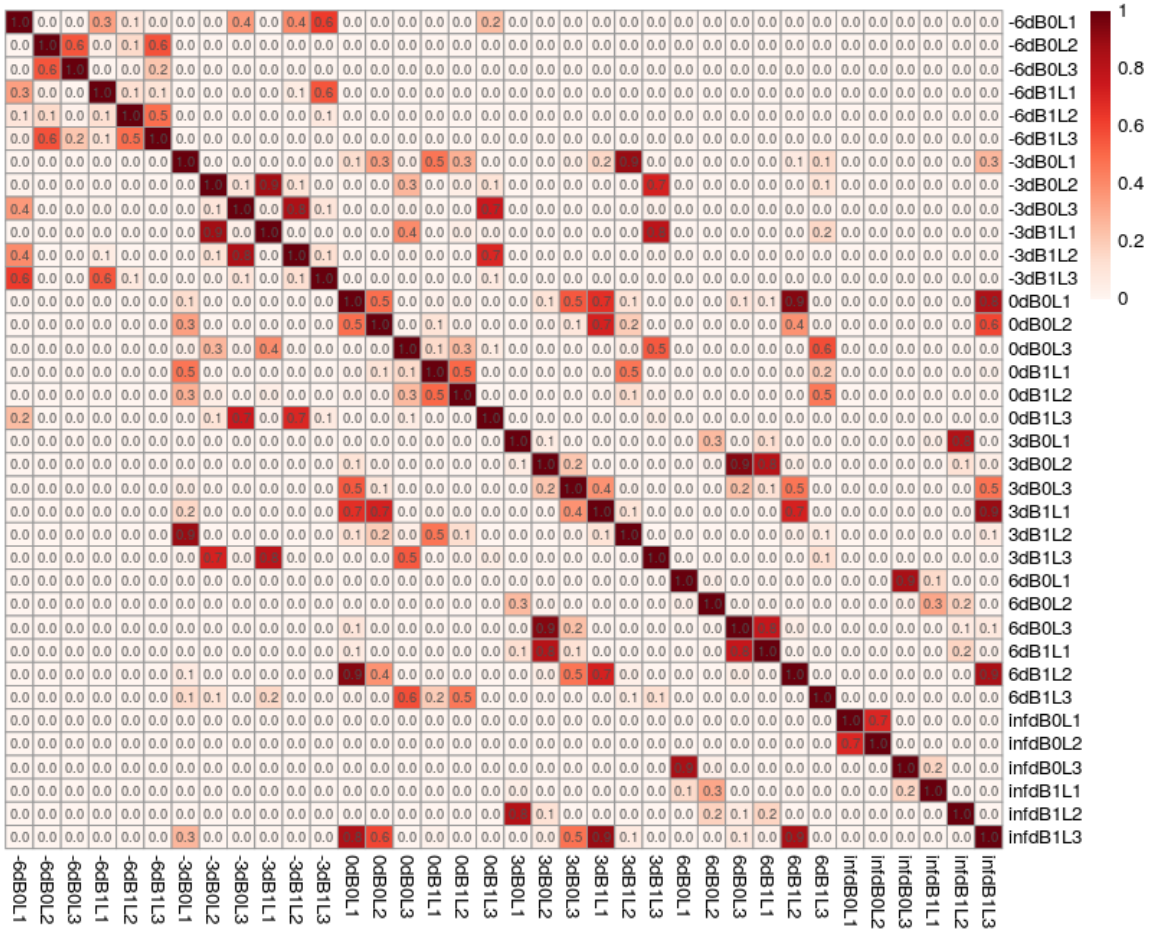
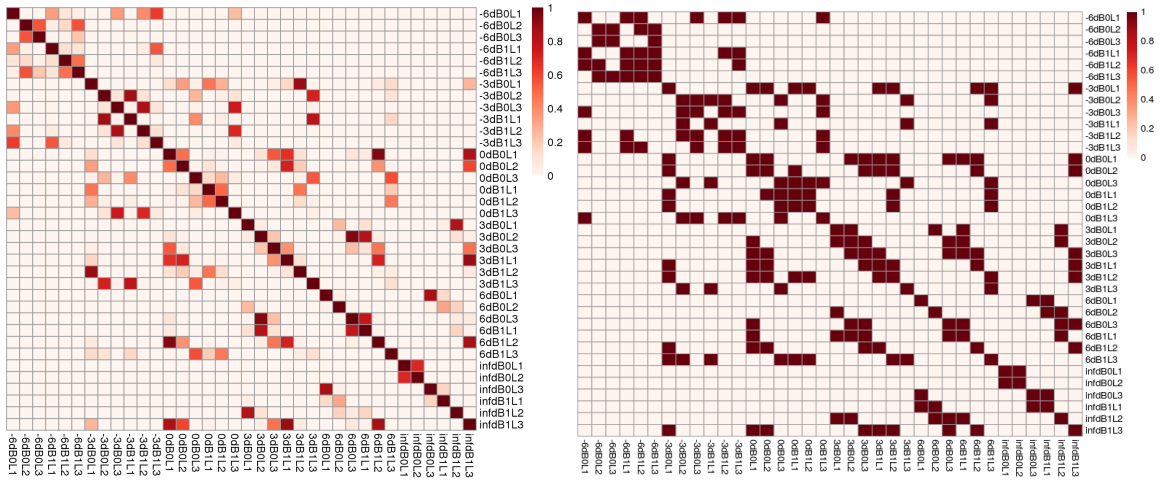


Fig. 3.6 *P*-matrix with *p*-values

all those pairs significantly differ from each other. From Figure 3.6, it can be concluded that all the experimental conditions with the lowest SNR (-6dB, top left corner) are close to each other and mostly differ from experimental conditions of the different noise level. Figure 3.7a shows the *P*-matrix again for comparison and Figure 3.7b shows the corresponding binary *P*-matrix, produced by threshold value 0.05 on *pvalues*. The binary *P*-matrix reflects the experimental groups which are significantly different from each other with 95% confidence.

To discover more structural information in the experimental groups, the binary P-matrix was clustered in a hierarchical tree. Figure 3.8 illustrates the hierarchical tree obtained from clustering and the corresponding rearranged *P*-matrix is shown in Figure 3.9. Given a branching point in the hierarchical tree, a cluster is defined as the collection of all the experimental conditions below the branching point. For instance, the branching point *R* defines a cluster consisting of two groups, namely *infdB0L1* and *infdB0L2*, and both groups are close in the sense that the impact on the attention of changing the experimental conditions



(a) *P*-matrix (b) Binary *P*-matrix for $p < 0.05$

Fig. 3.7 *P*-matrix

from *infdB0L1* to *infdB0L2* is smaller than changing to an experimental condition outside the cluster *R*.

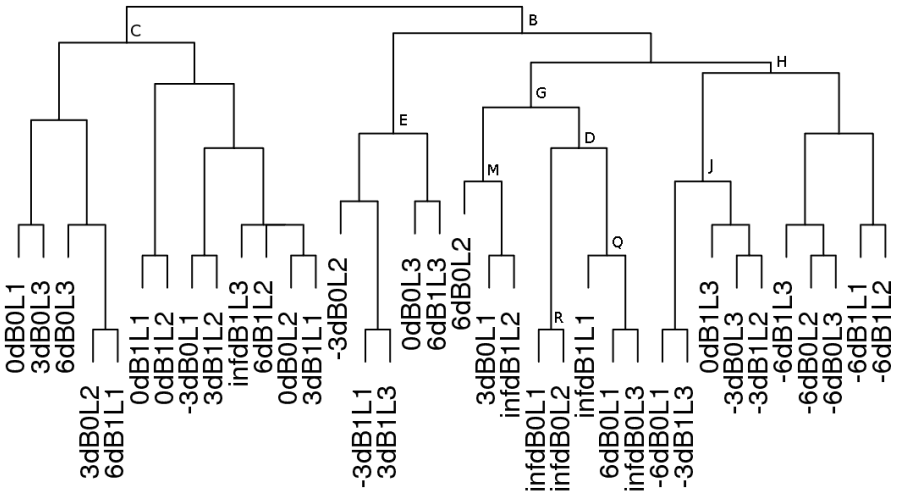


Fig. 3.8 Hierarchical clustering of experimental conditions obtained from binary *P*-matrix with threshold value of 0.05, $p < 0.05$.

The analysis of the hierarchical tree reveals interesting relationships about the experimental conditions in our study. By looking at cluster *Q*, it can be concluded that long semantic sentences in a noiseless environment produce the same effect as small semantic sentences in a noisy environment and small non-semantic sentences in a noiseless environment. Further-

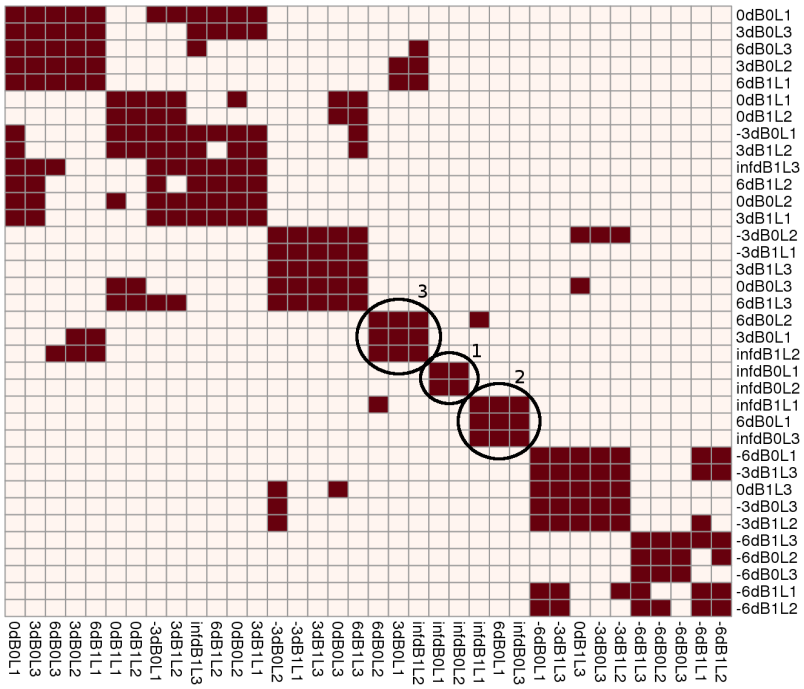
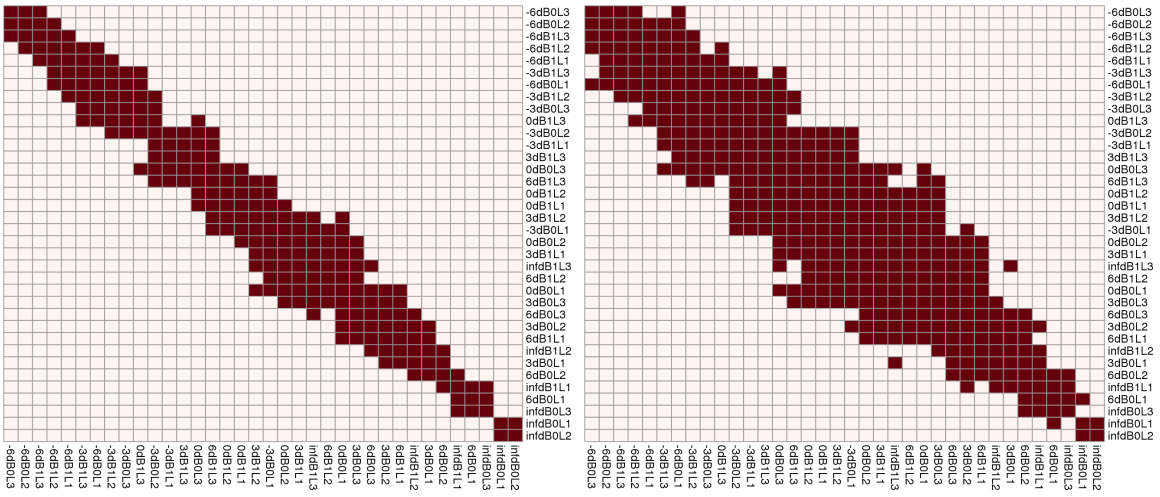


Fig. 3.9 Hierarchically clustered binary P-matrix with $p < 0.05$, where ■ represents $p \geq 0.05$ and □ represents $p < 0.05$.

more, the relationship between clusters R, Q, D and M suggest that the impact on attention of small semantic sentences when SNR = 3dB, is similar to non-semantic sentences of medium length in noiseless environments and semantic sentences of medium length in low-noise (6dB) environments. Cluster G includes the experimental conditions that perform the best. The mean attention score of these groups was above $A_{p,k\in G} > 55$. By contrast, cluster H includes the experimental conditions with the lowest mean attention score, namely below $A_{p,k\in H} < 17$. Interestingly, clusters H and C appear far from each other in the hierarchy and mostly differ in the noise level. Cluster H includes almost all the experimental conditions with low SNR, whereas cluster C includes experimental groups with high SNR and their mean attention score ranges from $27 < A_{p,k\in C} < 51$. Overall, the hierarchical tree can be decomposed into four major clusters, C, E, G, and H, within which the attention score is homogeneous.

The binary P-matrix shown in Figure 3.9 identifies which experimental conditions are significantly different from each other with a 95% confidence. The three groups labeled by 1, 2 and 3 in the rearranged P-matrix shown in Figure 3.9 correspond to clusters R, Q and M in Figure 3.8, respectively. The analysis of the first group corresponding to cluster R suggests that in noiseless environments (∞ dB SNR) small and medium lengths (L1 and L2) of semantic sentences are recognized with the same level of attention, whereas if the length



(a) $p < 0.05$ (b) $p < 0.001$

Fig. 3.10 Binary *P*-matrix, ranked with experimental conditions

of the sentence increases, non-semantic sentences are used or even low noise is introduced in the environment, the attention level drops significantly with 95% confidence. Following from this, another observation can be drawn by analyzing the cluster 2 identified in Figure 3.9, which corresponds to experimental groups *infdB0L3, 6dB0L1* and *infdB1L1*, namely in noiseless environments, long semantic sentences produce the same attention level as short non-semantic sentences and small semantic sentences with low noise. Another viewpoint of *P*-matrix is shown in Figure 3.10. Figure 3.10a and 3.10b are binary *P*-matrices with $p < 0.05$ and $p < 0.001$ respectively. The experimental groups are ranked by their overall mean attention score A_k. The bottom-right and top-left corners of *P*-matrix are corresponding to the experimental groups with the highest and lowest mean attention scores respectively.

3.2.3 Individual's auditory skill

For investigating the individual's auditory skill, the overall attention score of participants, A_p and experimental conditions A_k, was analysed. The overall attention score for individual varied from $11.16 < A_p < 65.86$. This variation could be explained by an individual's language or auditory skills. Figure 3.11 shows the individual differences as a heatmap, with overall attention score of each participant in each experimental condition $A_{p,k}$. In Figure 3.11, participants and experimental conditions are ranked as per their overall attention scores. The top participant (#1) has the highest overall attention score and it can be seen that this participant's performance is high in all the experimental conditions. In contrast, the participant with the lowest overall attention score (#19) exhibits the poorest performance in

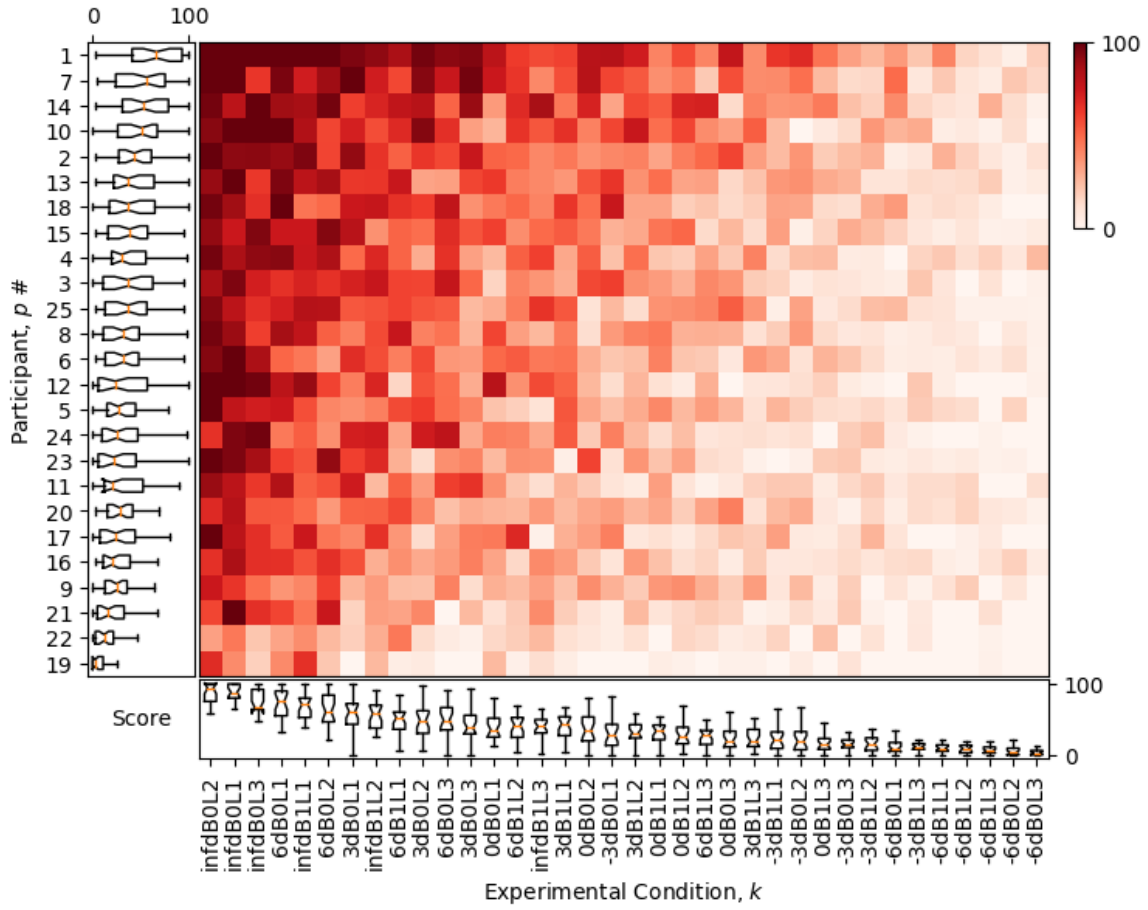


Fig. 3.11 Average attention score $A_{p,k}$, where the p (participant) and k (experimental condition) axes are arranged in descending order of computed A_p and A_k values, respectively.

almost all the experimental conditions. In general, it is apparent that there are considerable differences between participants, which can be accounted for based on their individual language competence.

Figure 3.12 shows the average attention score of each participant for each independent variable; noise level, length of stimulus and semanticity. The top graph shows the average attention score of each participant for the semantic and the non-semantic groups of stimuli. The middle graph shows the different lengths, L1, L2 and L3, and the bottom graph shows the different noise levels as labeled. Along with average attention score, the standard error is represented by a shaded bar. In principle, if shaded areas (standard errors) are not overlapping, both experimental conditions are significantly different from each other. It can be noticed that even though an individual's attention score varies considerably, almost all the participants exhibit significant difference for semanticity, as standard errors are not overlapping, except for participant 14 and 22. Similarly, for the length of stimulus, L1 and L3 have a significant

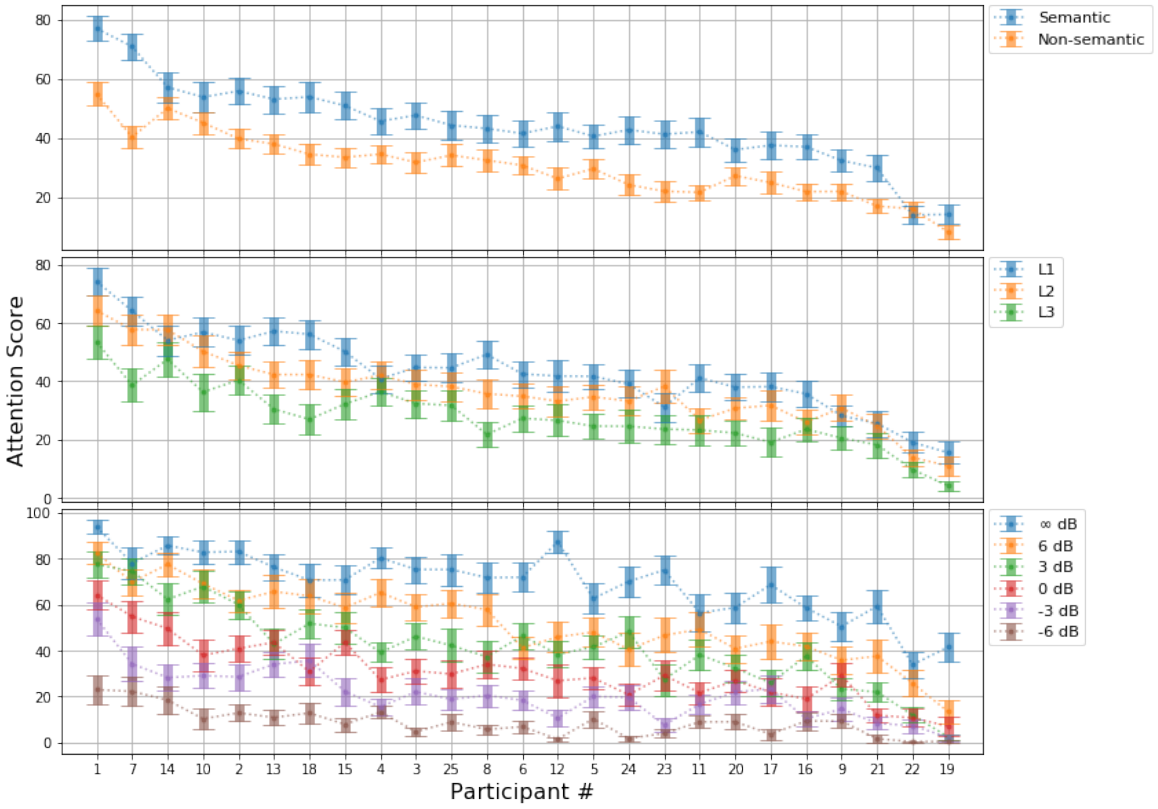


Fig. 3.12 Individual differences of participants for each independent variable: Noise level, length of stimulus, semanticity

effect on attention score for most of the participants, however, for L2, for some participants, it overlaps with L1 and for other, it overlaps with L3, except for participants 18 and 8, where L2 does not overlap with L1 or L3. For noise level, it is apparent, that the higher the difference between noise level, the more likely it is to have a significant effect.

In summary, the effects of auditory conditions, namely noise level and length of an auditory message have a significant impact on auditory attention, specifically for non-native speakers. While designing any environment for non-native speakers, these factors should be carefully considered. The examples of environments where auditory attention play a significant role in learning include online-classroom, training, serious games etc. Our results suggest that reducing the length of sentences always improves the level of auditory attention of non-native speakers. Background noise also has a very large impact on auditory attention. The semanticity also has a considerable impact on auditory attention, which suggests the use of simple and accessible vocabulary, without compromising the meaning, especially when non-native speakers are involved.

Chapter 4

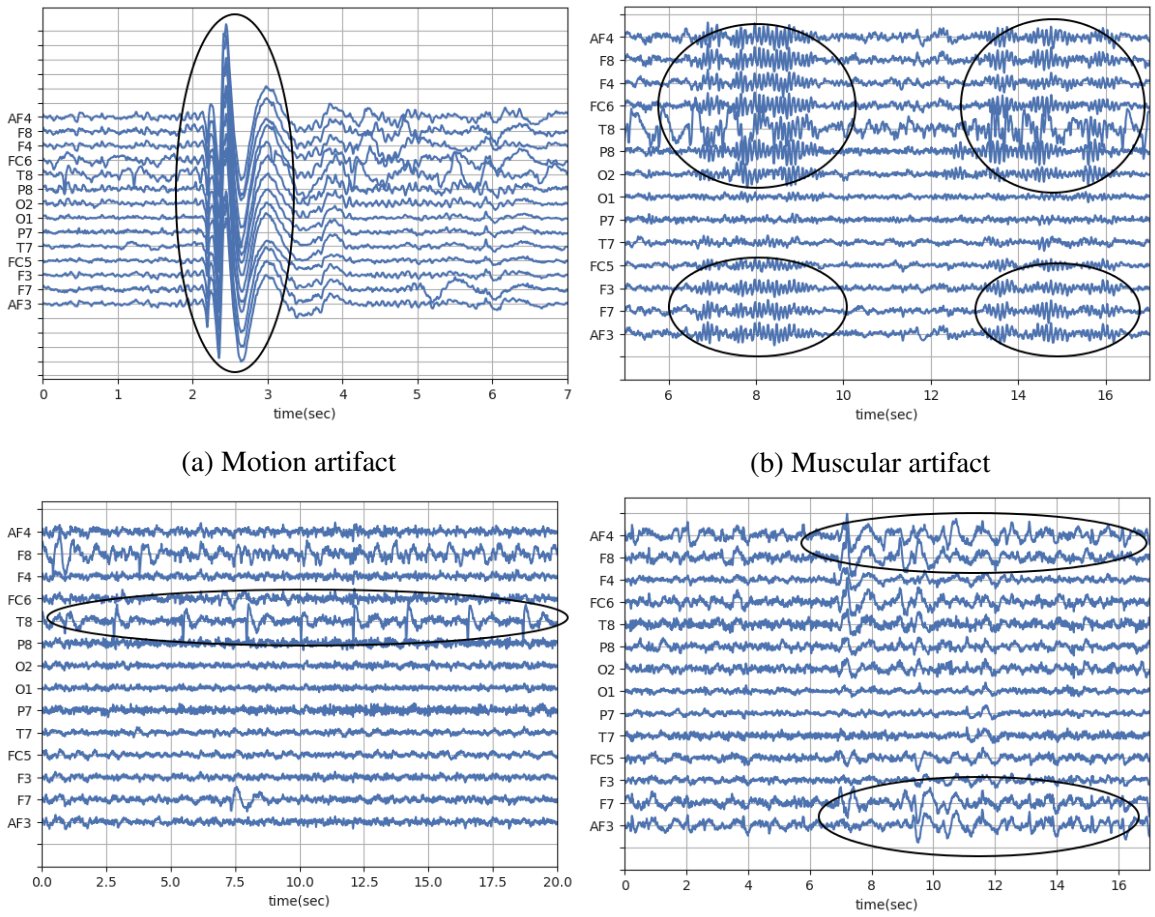
Wavelet based artifact removal algorithm

EEG signals are frequently contaminated by artifacts, external signals that are not caused by brain activity. Artifact removal algorithms have been proposed in the past to pre-process the EEG signals. However, two challenges still remain, namely the risk of removing useful information about the brain activity along with the artifact and the need for manually identifying artifactual components. In this chapter, we propose an algorithm based on Wavelet Packet Decomposition (WPD). The propose method allows us to control the degree of suppression of presumed artifacts. We investigate the performance of our WPD algorithm on the dataset created for auditory attention analysis without human intervention and compare it against Independent Component Analysis (ICA) based approaches.

4.1 Artifacts in EEG

EEG signals have been used extensively in many areas, such as neuroscience, psychophysiological research, cognitive science, neurolinguistics and many more. One of the main uses of EEG is to investigate neurological disorders in clinical studies. Neurological disorders include sleep disorder, epilepsy and othe neurological dysfunctions [49–51]. Due to technological development and ease of recording EEG with wearable devices, researchers have gone beyond the clinical studies to day-to-day activities and build BCI systems for various applications. In experimental settings, however, recorded of EEG signals are frequently contaminated with various artifacts. The most common types of artifacts are motion, ocular, muscular and cardiac artifacts [52], which are illustrated in Figure 4.1. Motion artifacts are caused by the physical movement of the subject's body. As shown in Figure 5.5a, motion artifacts produce a sudden high-valued spike in all the channels of an EEG recording. The muscular artifacts shown in Figure 5.5b are caused by any muscular contraction such as the muscular contraction produced by teeth grinding. Muscular artifacts produce high-frequency

bursts in EEG recordings as shown in Figure 5.5b. The cardiac artifacts shown in Figure 4.1c are caused by the electrical activity of the heart. They appear as a weak form of QRS cardiac wave and are most likely to appear in the channels near to ears (temporal lobe), though they can sometimes appear in the channels associated to the frontal lobe [53]. Ocular artifacts are slow oscillating waves appear on the frontal lobe, caused by the eye movements, as shown in Figure 4.1d.



(a) Motion artifact

(b) Muscular artifact

(c) Cardiac artifact

(d) Ocular artifact

Fig. 4.1 Common type of artifacts in EEG. Corresponding artifacts are circled in the figure.

Artifacts corrupt the EEG recording and may lead to misinterpretations during EEG analysis [54]. Even though there are many algorithms to remove artifacts, the risk of removing useful brain activity along with the artifacts that are meant to be eliminated remains high. Therefore, participants in traditional EEG studies are instructed to adopt a so-called minimal behavior [55], by not moving any body part unnecessarily, as this might cause artifacts. This, however, hinders the ability to study human cognition processes in general settings [56].

4.2 State-of-the-art algorithms

There is a considerable literature of algorithms to remove artifacts from EEG signals [57, 58]. Other physiological responses including ECG and PPG are also sensitive to artifacts [59, 60]. Early methods of detecting and removing artifacts from EEG include statistical approaches with interpolation [61] and regression [62]. The state-of-the-art algorithms for artifact removal are based on Blind Source Separation (BSS) using ICA. Over the decade, ICA has been improved by incorporating supervised classification [63, 64], statistical measures (z-scores) [65], wavelet denoising [66, 67], and Second Order Blind Identification (SOBI) [68]. In addition, mathematicians have explored different methods to compute independent components with different assumptions to improve robustness and efficiency. Among other methods, it is worth mentioning FastICA, InfoMax [69, 70], and Extended-Infomax [71] which were used in [55] for removing motion artifacts from high-density EEG recordings during walking and running by using a template of artifacts and combining it with regression techniques.

Most of these approaches require an expert to manually select the artifactual components and other approaches use additional reference signals (e.g. EOG, EMG, ECG) [72], carrying the potential information about artifacts present in EEG signals. Many of the proposed algorithms are tested on simulated artifacts or with reference signals, so as to evaluate the performance of the algorithm. However, in a practical setting where neither ground truth of artifacts nor reference signals are available, it can not be assessed whether the applied algorithm removed any useful information along with the artifact. Therefore in this chapter, we proposed an algorithm based on Wavelet Packet Decomposition (WPD), that allows controlling the suppression or removal of presumed artifacts. Wavelet-based approaches allow describing time-localized events and thus are well-suited for identifying the artifacts localized in EEG signals. In addition, the proposed algorithm defines a few parameters that can be tuned to control the type of artifact suppression and improve the performance in subsequent stages, such as predictive stages.

4.3 Artifact removal method

4.3.1 Wavelet packet decomposition

In WPD, signals are decomposed onto a wavelet packet basis at different scales [73] [74]. A wavelet packet basis for j-level decomposition is defined as a collection of signals $\{\psi_j^i(n-2^j k)\}_{k\in\mathbb{Z}}$ where $i \in \mathbb{Z}^+$, $0 \leq i \leq 2^j - 1$. The wavelet packet bases $\psi_j^i(n)$, are generated recursively from the so-called scaling and wavelet functions, $\psi_1^0(n) = \phi(n)$ and $\psi_1^1(n) = \psi(n)$

respectively, as follows:

$$\psi_j^{2i}(n) = \sum_k h(k)\psi_{j-1}^i(n-2^{j-1}k) \tag{4.1}$$

$$\psi_j^{2i+1}(n) = \sum_k g(k)\psi_{j-1}^i(n-2^{j-1}k) \tag{4.2}$$

where $h(n)$ and $g(n)$ are lowpass and highpass quadrature mirror filters respectively [73, 75] and defined as;

$$h(k) = \langle \psi_j^{2i}(u), \psi_{j-1}^i(u-2^{j-1}k)\rangle \tag{4.3}$$

$$g(k) = \langle \psi_j^{2i+1}(u), \psi_{j-1}^i(u-2^{j-1}k)\rangle \tag{4.4}$$

The decomposition of a signal $x(n)$ onto the wavelet basis $\psi_j^i(n)$ at level j can be written as

$$x(n) = \sum_{i,k} X_j^i(k)\psi_j^i(n-2^j k) \tag{4.5}$$

where $X_j^i(k)$ is the k^{th} wavelet coefficient of i^{th} packet, at level j. The wavelet coefficient $X_j^i(k)$ describes the intensity of the localized wavelet $\psi_j^i(n-2^j k)$ and is defined by the projection

$$X_j^i(k) = \langle x(n), \psi_j^i(n-2^j k)\rangle \tag{4.6}$$

Figure 4.2 shows the 4-level WPD of signal $x(n)$, where LP and HP are lowpass and highpass filters with impulses responses $h(n)$ and $g(n)$, respectively, followed by a downsampling stage by a factor of two.

Let $x(n)$ represent a recorded EEG signal. By using bioelectric principles, $x(n)$ can be modeled as the sum of a source signal $s(n)$ induced by the brain activity (cerebral activity) and an aritifact signal $v(n)$,

$$x(n) = s(n) + v(n) \tag{4.7}$$

The source signal $s(n)$ is commonly assumed to be normally distributed for short duration with zero mean, $s(n) \sim N(0,\sigma)$, where σ^2 denotes the variance of $s(n)$ [52]. By contrast, it is generally assumed that the artifact signal $v(n)$ is temporally localized, not normally distributed and its variance is locally high.

Due to the linear nature of the transformation described by (4.6), the wavelet coefficients of $x(n)$, $X_j^i(k)$, can be expressed as the sum of the wavelet coefficients $S_j^i(k)$ and $V_j^i(k)$ of, respectively, $s(n)$ and $v(n)$,

$$X_j^i(k) = S_j^i(k) + V_j^i(k) \tag{4.8}$$

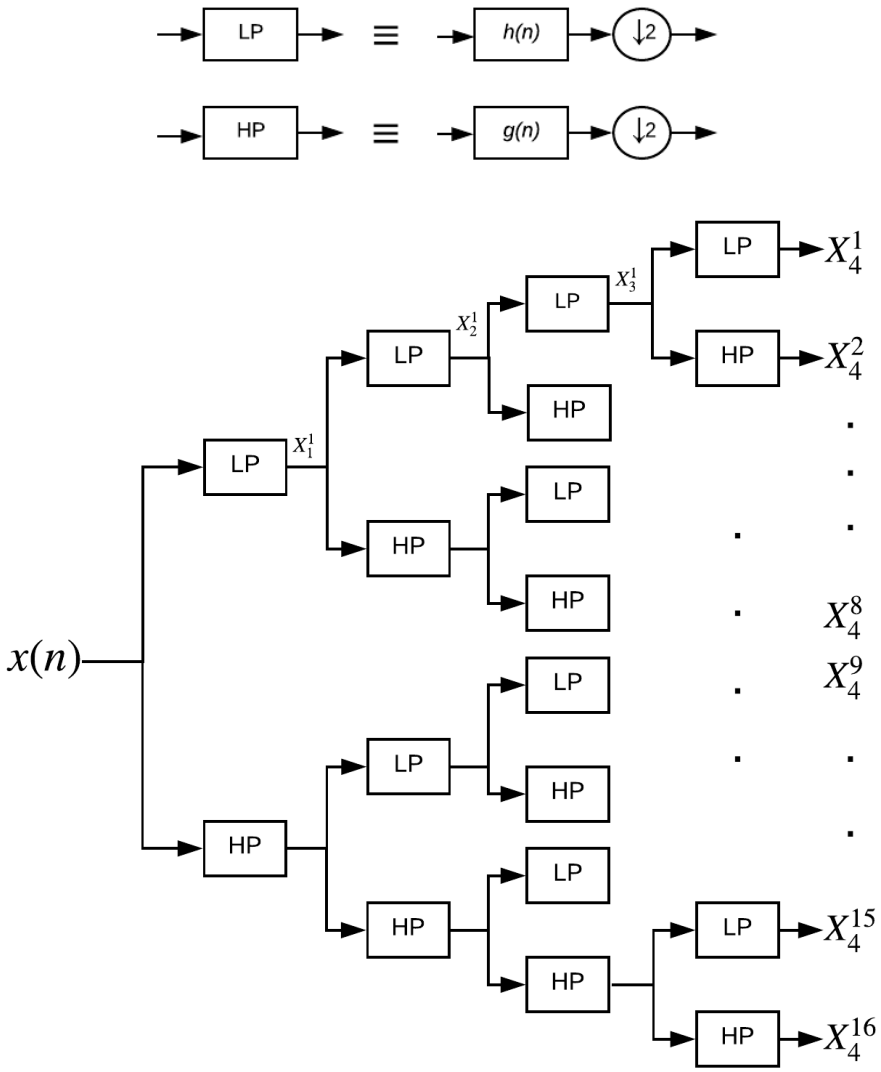


Fig. 4.2 Wavelet Packet decomposition, 4-levels. LP and HP are lowpass and highpass filters followed by decimation by factor of 2. LP and HP are associated with scaling and wavelet function.

Since the variance of $v(n)$ is locally high, its wavelet coefficients $V_j^i(k)$ will be sparse and non-zero coefficients will have a high magnitude. This will result in a local increase of the variance of the recorded signal $x(n)$, which will manifest in several coefficients $X_j^i(k)$ being comparatively high. This observation about the behaviour of the wavelet coefficients of contaminated EEG signals is the basis of our proposed approach for removing artifacts. The WPD of one second, single channel EEG signal with Daubechies ($db3$) is shown in Figure 4.3. It can be observed that the signal is decomposed into localized wavelet components of different scales at each packet. For example, the wavelet components of packet-1; Xr^1

are wider, in contrast to packet-16; Xr^{16}. For removing artifactual components, the proper selection and removing such wavelet components, which are more likely to represent the artifacts, are the keys to the proposed algorithm.

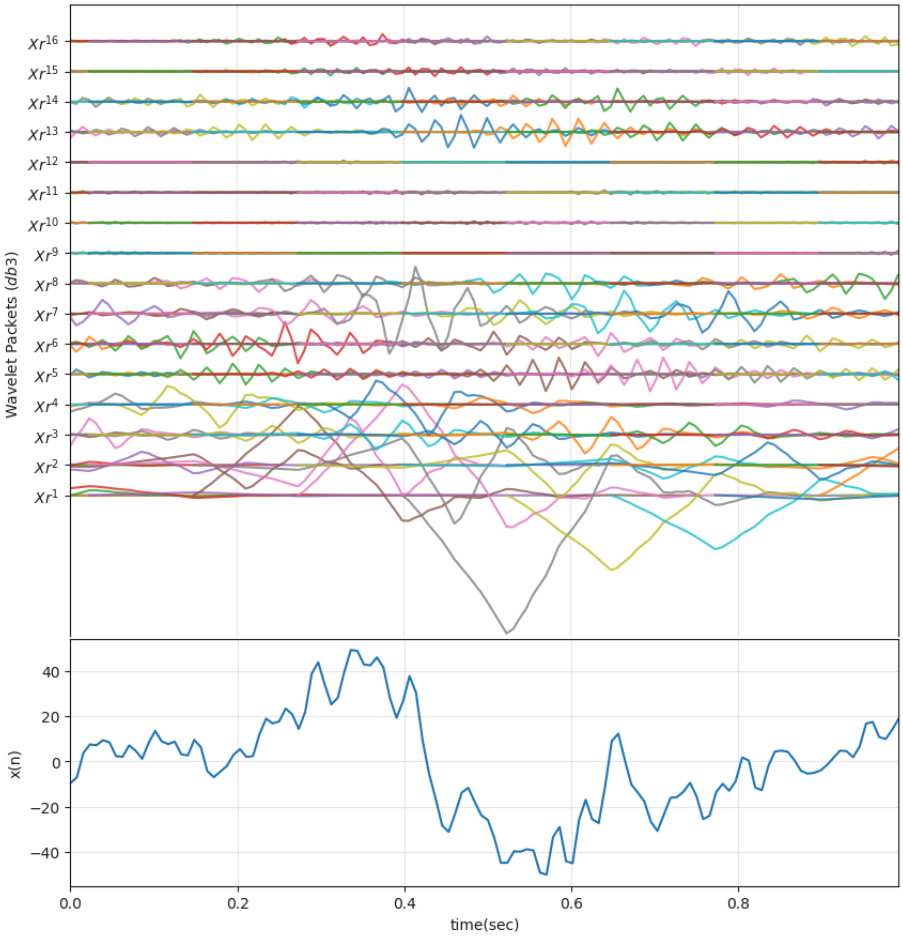


Fig. 4.3 4-Level wavelet packet decomposition of 1 sec $x(n)$ using $db3$

4.3.2 Filtering method

By assuming that artifacts in EEG signal are localized and hence are represented by a restricted number of wavelets, we propose an approach to adjust the attenuation level of the presumed artifacts and limit the distortion produced on the source signal of interest. Our approach defines a filtering function $\lambda(\cdot)$, operating in the wavelet-domain that produces the reconstructed signal $\tilde{x}(n)$:

$$\tilde{x}(n) = \sum_{i,k} \lambda\left(X_j^i(k)\right) \psi_j^i(n - 2^j k) \tag{4.9}$$

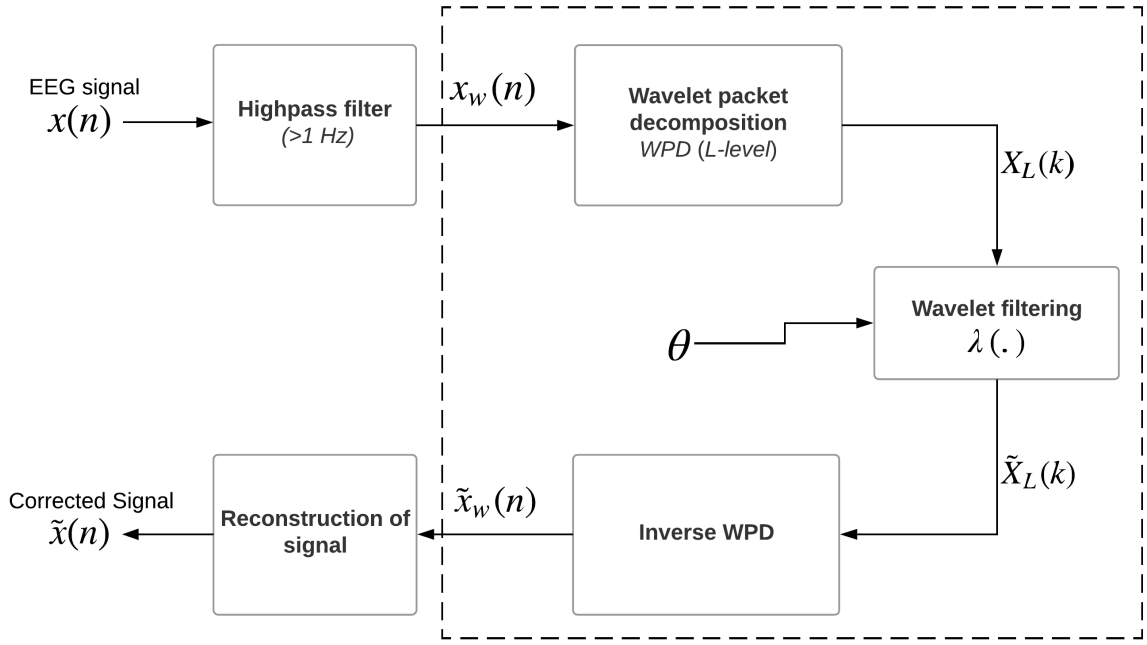


Fig. 4.4 Block diagram of the proposed wavelet filtering method.

The block diagram of the proposed approach is shown in Figure 4.4. First, the EEG signal $x(n)$ passes through a high-pass filter with a cut-off frequency of 1 Hz. Then, segments of size N are extracted from the filtered signal with 50% overlap and decomposed onto the L-level wavelet basis, resulting in coefficients $X_L^i(k)$. We will denote the set of coefficients $X_L^i(k)$ by $X_L(k)$,

$$X_L(k) = \begin{bmatrix} X_L^0(k) \\ X_L^1(k) \\ \vdots \\ X_L^{2^L-1}(k) \end{bmatrix}. \tag{4.10}$$

The wavelet coefficients $X_L(k)$ are subsequently filtered by the wavelet filtering function $\lambda(\cdot)$ with tuning parameter θ, resulting $\tilde{X}_L(k)$. A segment of the signal $x_w(n)$ is reconstructed by inverse WPD (4.9) resulting $\tilde{x}_w(n)$. Finally the entire signal $\tilde{x}(n)$ is synthesized from all the reconstructed segments with overlapping add method.

As a consequence of the assumption that the artifacts produce high variance locally in the EEG signal, a high variance in some wavelet coefficients $X_j^i(k)$ can be expected [76]. Since each wavelet coefficient is a localized representation of the decomposed signal, coefficients causing high variance can be attenuated or removed by setting an appropriate

threshold. This is the objective of wavelet filtering function $\lambda(\cdot)$. However, according to (4.8) in general, every wavelet coefficient will also contain information about the source signal $s(n)$. Therefore a complete elimination of such wavelet components might eliminate useful information. To deal with this issue, in the following subsection we present three different implementations of the wavelet filter $\lambda(\cdot)$, namely Elimination $\lambda_e(\cdot)$, Linear attenuation $\lambda_a(\cdot)$, and Soft-thresholding $\lambda_s(\cdot)$.

4.3.3 Implementations of the wavelet filter

Elimination

The simplest option for wavelet filtering is to completely eliminate the wavelet components, which exceed a given threshold θ_α. This approach, therefore, assumes that such wavelet components are purely artifactual and contain no useful information (i.e. $Pr(max|S(k)_j| > \theta_\alpha) = 0$). The elimination wavelet filter $\lambda_e(w)$ is defined as follows,

$$\lambda_e(w) = \begin{cases} w & \text{if } |w| \leq \theta_\alpha \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where w is a wavelet coefficient.

Linear Attenuation

As we discussed, due to their additive nature, each artifactual wavelet component also contains information from the cerebral activity of the brain. Therefore, we assume that the higher the amplitude of a wavelet component above the threshold θ_α, the lower the probability that it contains useful information about the source signal. In this mode, wavelet components are preserved if their amplitude is low, attenuated linearly beyond a predefined threshold level θ_α, and eliminated beyond another level θ_β. The wavelet filter in linear attenuation mode $\lambda_a(\cdot)$ is defined as follow;

$$\lambda_a(w) = \begin{cases} w & \text{if } |w| \leq \theta_\alpha \\ sgn(w)\theta_\alpha \left(1 - \frac{|w|-\theta_\alpha}{\theta_\beta - \theta_\alpha}\right) & \text{if } \theta_\alpha < |w| \leq \theta_\beta \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

where $sgn(\cdot)$ is the signum function.

Soft-Thresholding

The objective of soft-thresholding is not to completely eliminate but to attenuate unusually strong wavelet components to a predefined level θ_α, so that the distortion caused to the underlying source signal can be controlled. In choosing the level θ_α, we are assuming that the probability that the wavelet coefficients of the source signal are higher than θ_α, is low. By limiting the magnitude of the wavelet components, we seek to reduce the effect of artifacts on the signal while preserving the features of the source signal captured by them. The wavelet filter in soft-thresholding mode attenuates the wavelet coefficients greater than θ_γ and smoothly limit them to θ_α with hyperbolic tangent function. The soft-thresholding mode $\lambda_s(\cdot)$ is defined as

$$\lambda_s(w) = \begin{cases} w & \text{if } |w| < \theta_\gamma \\ \frac{1-e^{-\alpha w}}{1+e^{-\alpha w}}\theta_\alpha & \text{otherwise} \end{cases} \tag{4.13}$$

The parameter α is defined as

$$\alpha = -\frac{1}{\theta_\gamma}\log\frac{\theta_\alpha - \theta_\gamma}{\theta_\alpha + \theta_\gamma}. \tag{4.14}$$

where $\theta_\alpha > \theta_\gamma$. The soft-thresholding filter implements a soft transition from a linear behaviour to a saturating one. Parameters θ_α and θ_γ determine the speed of the transition, where $\theta_\alpha \approx \theta_\gamma$ corresponding to a hard-threshold. In the following sections, we will consider a ratio $\theta_\gamma = 0.8\theta_\alpha$.

Figure 4.5 shows the characteristics of wavelet filter in three operating modes, namely elimination $\lambda_e(\cdot)$, linear attenuation $\lambda_a(\cdot)$ and soft-thresholding $\lambda_s(\cdot)$ for threshold $\theta_\alpha = 200$ and $\theta_\gamma = 0.8\theta_\alpha = 160$ and $\theta_\beta = 2\theta_\alpha = 400$.

4.3.4 Threshold selection

Wavelet transform has been used for denoising the signal from white-gaussian noise and optimal threshold as defined in [77] is $\hat{\sigma}\sqrt{2logN}$, where N is the length of signal and for wavelet coefficients w, $\hat{\sigma} = median(|w|)/0.6745$ is the estimate of noise variance. Any wavelet coefficient below the threshold is estimated to zero (set to zero) to recover the noise-free signal. However, unlike the denoising method, our assumption is that the artifacts are not white-gaussian processes, rather the signals of interest are normally distributed with zero mean i.e. $s(n) \sim N(0,\sigma)$ for short duration [52]. Our approach is to select the threshold, such that the wavelet coefficient of source signal lies below the selected threshold.

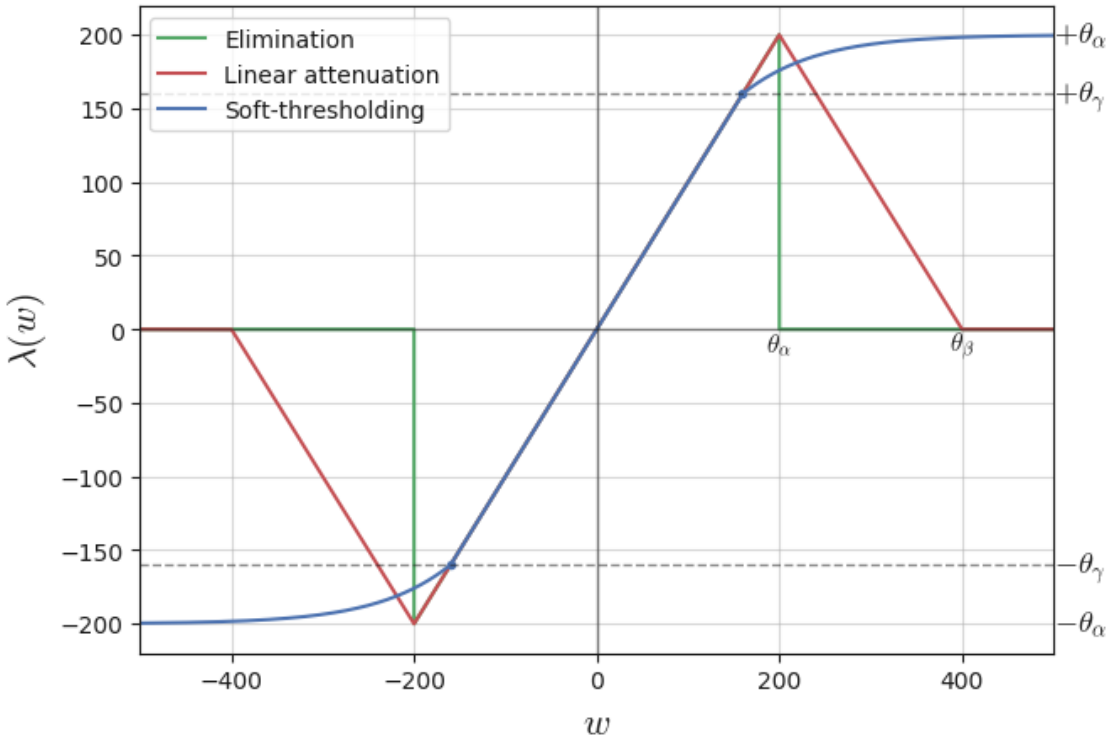


Fig. 4.5 Characteristics of the wavelet filter $\lambda(.)$ for different operating modes, Elimination, Linear Attenuation, and Soft-thresholding.

One approach for selecting the threshold θ_α is to choose it manually by inspecting the recorded EEG signal carefully and identifying the highest value of wavelet coefficients from artifact-free segment of signal, such that the probability of the magnitude of any wavelet coefficient of artifact-free signal being greater than threshold, θ_α is near to zero, i.e. $Pr(max|S_j| > \theta_\alpha) \sim 0$. However manual selection is not an efficient method, as it needs an adequate understanding of the statistical properties of the recorded EEG signal.

An efficient approach to select the threshold θ_α is to formulate it based on the statistics of the signal for a given short duration. In line with our assumption that artifacts cause a high variance in the EEG signal along with outliers, we select the threshold θ_α based on the variability of the recorded EEG signal. As variance is sensitive to outliers, while Interquartile-range (IQR) is robust against outliers, IQR is used to select the threshold to avoid the high difference in threshold values for consecutive segments of the signal. Since $\lambda(.)$ is applied to wavelet coefficients, the selection of threshold should be on the variance of wavelet coefficients rather than on EEG signal itself. The relationship between the variance of the signal and its wavelet coefficients is nondeterministic. However, there is a high correlation, as pointed out by [76, 78] that high values of wavelet variance reflect the presence of a greater

number of peaks, a greater magnitude of the signal, or both. In principle, low IQR of wavelet coefficients suggests less variance, thus lower probability of presence of assumed artifacts in the EEG signal. The impact of threshold θ_α on the reconstructed signal of different standard deviation (SD) was observed by computing the energy ratio (energy of reconstructed signal E_r over the energy of original signal E_x). The effect is shown in Figure 4.6a, which exhibits the exponential relationship, which tends to be linear for a signal with high SD. The exponential relationship leads us to formulate the threshold θ_α selection as follow;

$$\theta_\alpha = f_\beta(r) = \begin{cases} k_2 exp\left(-\beta \frac{100}{k_2} \frac{r}{2}\right) & \text{if } r \geq -\frac{2k_2}{100\beta} log(k_1/k_2) \\ k_1 & \text{else} \end{cases} \tag{4.15}$$

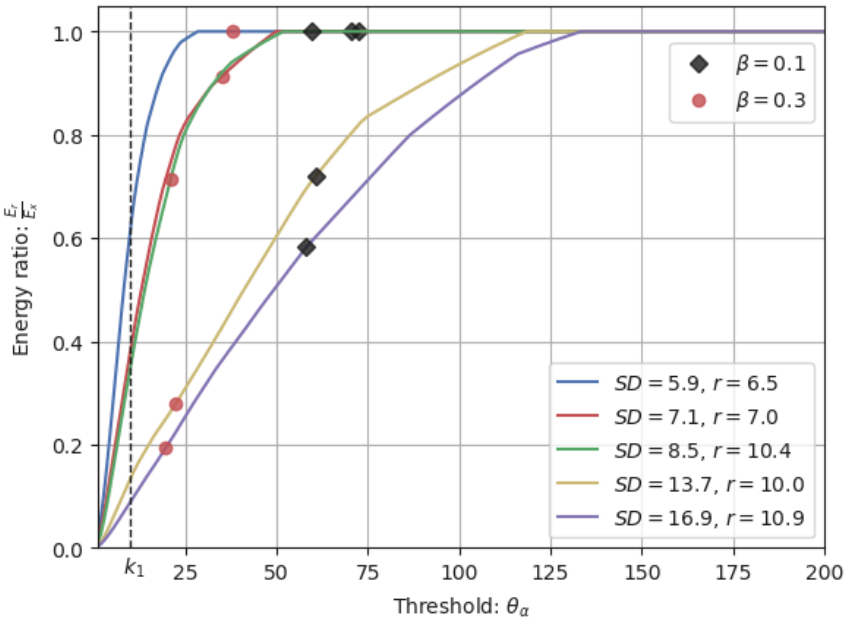
assuming;

$$Pr\Big(max|S_j| > f_\beta(r)\Big) \to 0, \quad r \to \infty$$

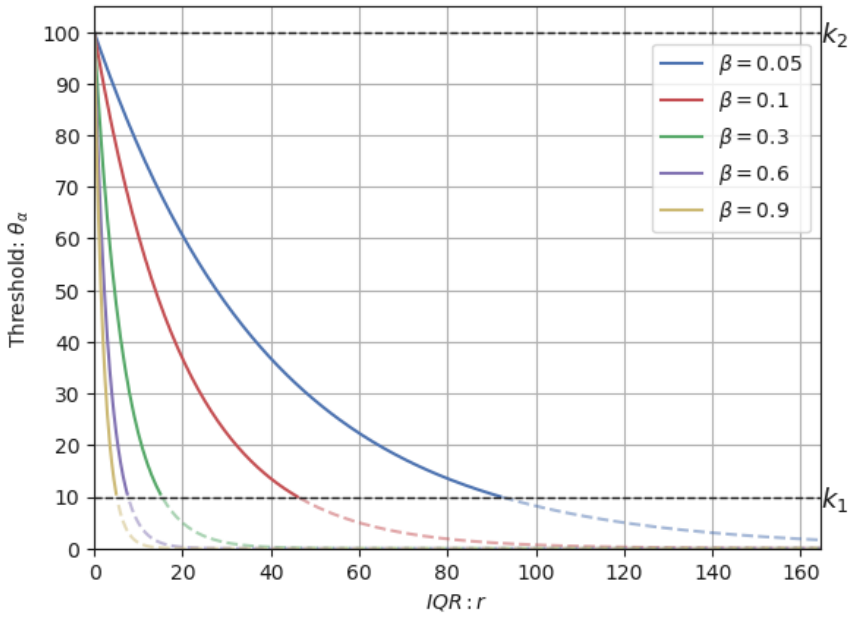
where r is the IQR of wavelet coefficients, β is the steepness (attenuation constant) and k_1 and k_2 are lower and upper bound on threshold θ_α respectively i.e. $\theta_\alpha \in [k_1, k_2]$. Figure 4.6b shows $\theta_\alpha = f_\beta(r)$ for $k_1 = 10$ and $k_2 = 100$. By definition, lower and upper bounds [k_1, k_2] on threshold θ_α sets the limit on threshold value. A lower bound k_1 prevents a given segment of the reconstructed signal to be zero, in case of very high IQR range. The upper bound k_2 and attenuation constant β control the curve for selecting the threshold θ_α.

The effect of the threshold θ_α on the signal depends on the variation of the signal values. Figure 4.6a shows an effect of threshold on five different segments of the signal taken from the EEG signal, varying the threshold from 0 to 200. As Figure 4.6a shows, these five segments have different SD and IQR r of respective wavelet coefficients. The energy conservation computed by the ratio of the energy of reconstructed signal E_r to the energy of the original signal E_x, shows how much original signal is affected. For example, an energy ration (E_r/E_x) equal to 1 indicates no effect of the threshold. It can be noticed that, for similar values of IQR which have different SDs, the computed threshold θ_α using equation (4.15) has a dramatic change of effect on the signal. For example, for IQR = 10.4, with SD = 8.5, any threshold above 50 does not affect the signal and the entire signal is preserved. On the other hand, for IQR = 10 and SD = 13.7, a threshold 50 reduced the energy of the original signal by 60%.

It can also be observed that for low SD in an input signal, increasing threshold θ_α quickly results in $Er/Ex = 1$, which means that wavelet filtering has no effect on a signal. Furthermore, the threshold θ_α computed by equation (4.15) with $\beta = 0.1$ does not affect signal with low SD and preserves the approximately 60% to 70% of energy of high SD input signal, while θ_α with $\beta = 0.3$ has little effect on low SD signal and suppresses high SD signal



(a) Energy ratio



(b) Threshold θ_α verses range IQR, r

Fig. 4.6 Threshold (a) Energy conservation of signal computed with E_r/E_x ratio for different threshold values θ_α for $\beta = 0.1$ and $\beta = 0.3$ (b) The curve of threshold selection equation (4.15) for different steepness value β and lower and upper bounds [10,100].

substantially. Since the steepness β is very sensitive for the curve, the alternative choice is to adjust the estimation of range r. Rather than fixing range r to be IQR (50%), it can be

adjusted to select 70% or 90% of middle range, which can be considered as the Interpercentile range (IPR). The proposed approach is encapsulated as a whole artifact removal procedure in Algorithm 1.

Algorithm 1 Tunable algorithm for artifact removal from EEG signal using wavelet decomposition

Input: Single channel EEG signal $x(n)$
Output: Corrected EEG signal $\tilde{x}(n)$
Parameters choice :
Wavelet family:(say *db3*), window size: N samples
Bounds on threshold θ_α: $[k_1, k_2]$ (say [10, 100] or [0.1, 1.0])
Wavelet filtering mode: $\lambda_e(\cdot)$, $\lambda_a(\cdot)$, or $\lambda_s(\cdot)$ and corresponding θ_β ($\theta_\beta = 2\theta_\alpha$) or θ_γ ($\theta_\gamma = 0.8\theta_\alpha$)
Threshold selection parameter: β (say 0.1), *IPR* (say 50%)
Procedure:
1: Filter the input signal $x(n)$ with high pass filter of cut-off frequency 1 Hz: $x_f(n) \leftarrow x(n)$
2: **while** all windows of $x_f(n)$ are extracted **do**
3: Extract a window of signal with 50% overlapping: $x_w(n) \leftarrow x_f(n)$
4: Compute L-level WPD: $X_L(k) \leftarrow WPD(x_w(n))$
5: Compute θ_α ▷ equation (4.15)
6: Apply wavelet filtering: $\tilde{X}_L(k) \leftarrow \lambda(X_L(k))$
7: Reconstruct signal with IWPD: $\tilde{x}_w(n) \leftarrow IWPD(\tilde{X}_L(k))$
8: Synthesize the entire signal with overlapping add method: $\tilde{x}(n) \leftarrow \tilde{x}_w(n)...$

4.4 Experiments

4.4.1 Dataset

We used the dataset collected for auditory attention introduced in Chapter 2. The dataset has 14-channels of EEG recordings with three subtasks labels of *Listening, Writing*, and *Resting*. All the listening segments are labeled with their corresponding noise level (SNR), semanticity and attention scores. We evaluate the performance of our algorithm on four predictive tasks explained in Section 2.7 as follow

1. ***LWR classification***: Prediction of the participant's task state, namely Listening, Writing, or Resting labeled as 0,1 and 2 respectively.

2. ***Semanticity classification***: Prediction of the semanticity of audio stimulus experienced by a participant in a listening task, labeled as 0 and 1 for semantic and non-semantic respectively.

3. ***Noise level prediction***: Prediction of the noise level in an audio stimulus experienced by participants during a listening task. Though we used six levels of noise in the experiment, we merged them to form three classes. The three classes are -6 dB, -3 dB to 3 dB and 6 dB to ∞ dB, labeled as 0, 1, and 2 respectively.

4. ***Attention score prediction***: Prediction of attention score computed for listening segments. As attention score ranges from 0 to 100, this is treated as a regression task.

4.4.2 Parameter choice for experiment

For removing artifacts from EEG signals, we followed the steps summarized in Algorithm 1. The choice of the wavelet for this experiment was a Daubechies wavelet, *db3*. The window size for processing the signal is set to $N = 128$. The bounds on threshold θ_α are set to [10, 100]. The input signal $x(n)$ is first filtered with a 5^{th} order IIR bandpass filter, with cut-off frequency of $1 - 40$ Hz. Overlapping windows of size N were extracted from filtered signal, $x_w(n) \in \mathbb{R}^{128}$. For input signal size of 128 and wavelet *db3*, the maximum levels of wavelet packet decomposition is $L = 4$, which results into 16 wavelet packets, with 12 coefficients each [79]. Each window $x_w(n)$ was decomposed with WPD resulting wavelet coefficients $X_4(k) \in \mathbb{R}^{192}$. The wavelet coefficients $X_4(k)$ are then filtered with $\lambda(.)$ with three operating modes, as defined by equations (4.11), (4.13), (4.12) with different choices of parameters β and *IPR*. The different values for β were chosen as 0.6 and 0.9 and for *IPR*, 50% and 70% were chosen. Each combination of parameter results into $\tilde{X}_4(k)$, which was used to reconstruct the signal $\tilde{x}_w(n)$ using inverse WPD. Once all the windows are processed, $\tilde{x}(n)$ is synthesized with the overlap add method.

4.4.3 Artifact removal with ICA

For the sake of comparison, we applied the ICA-based artifact removal algorithm on the same dataset. Since there are very few articles proposing the ICA-based algorithm that do not require experts to manually select artifactual components and additional reference signals of artifacts (e.g. EOG), we adapted an algorithm proposed in [80] and heuristics of kurtosis as explained in [65, 81]. The algorithms proposed in [80] suggest to remove eye blink artifacts that have similar characteristics as those we have assumed to be removed (i.e. cases with a high variance for short duration). However this algorithm is limited to eye blink artifacts only, so we added more constraints to it to remove additional artifacts producing high variance. In [80], the correlation based index (CBI) is computed as a correlation of independent components to the frontal lobe and identifying it as blinking artifact. Following

the same principle, we computed the CBI of independent components to all the channels and identifying components that are correlated to more than 80% of the channels, assuming them as motion artifacts. In addition, we computed the kurtosis of independent components and identified as artifacts those components with an absolute value of kurtosis greater than equal to 2. We used three popular methods to compute independent components, namely FastICA, InfoMax, and Extended-InfoMax.

4.5 Results and discussions

For comparing our algorithm with the ICA-based automatic algorithm, the outcomes of the above experiment are analyzed in four different ways. The first two approaches inspect the signal visually before and after applying the algorithm. Next, we compute the improvement in a correlation of spectral features and the target value of auditory tasks explained in Section 4.4.1. Finally, we measure the performance of all the four predictive tasks and compare it with the ICA-based approach. We also analyse the effects of parameter β on the predictive tasks.

4.5.1 Visual inspection

As most of the artifacts in the EEG signal are visually detectable, the first approach consists of analysing the performance visually. A 10-second segment of the 14-channel filtered EEG signal and its corresponding corrected segments with ICA-based approach and WPA-based proposed approach are shown in figures 4.7 and 4.8, respectively. As explained, ICA-based approach with three different methods of computing ICA, namely FastICA, InfoMax, and Extended-InfoMax are used. For WPA-based approach, wavelet filtering modes Soft-thresholding, Linear attenuation and Elimination with $\beta = 0.6$, $IPR = 50$ are used.

Three main different kinds of artifacts are illustrated in the filtered signal in Figure 4.7. The first one is similar to muscular artifacts and can be identified by high frequency present in the first seven channels in the first two seconds (253-255 sec). The second kind of artifact appears to be a motion artifact (a peaky impulse present in between 256-257 and 258-259 seconds). The third kind corresponds to blinking artifacts, a slow oscillation present in the FC5 channel in the last 3 seconds. It can be observed from Figure 4.7 that, FastICA, InfoMax, and Extended-InfoMax are able to remove peaky artifacts, although FastICA and InfoMax are left with a small distinct peak in all the channels. With FastICA, slow oscillating artifacts are completely removed, but the segment between 255-257 seconds shows that the neural

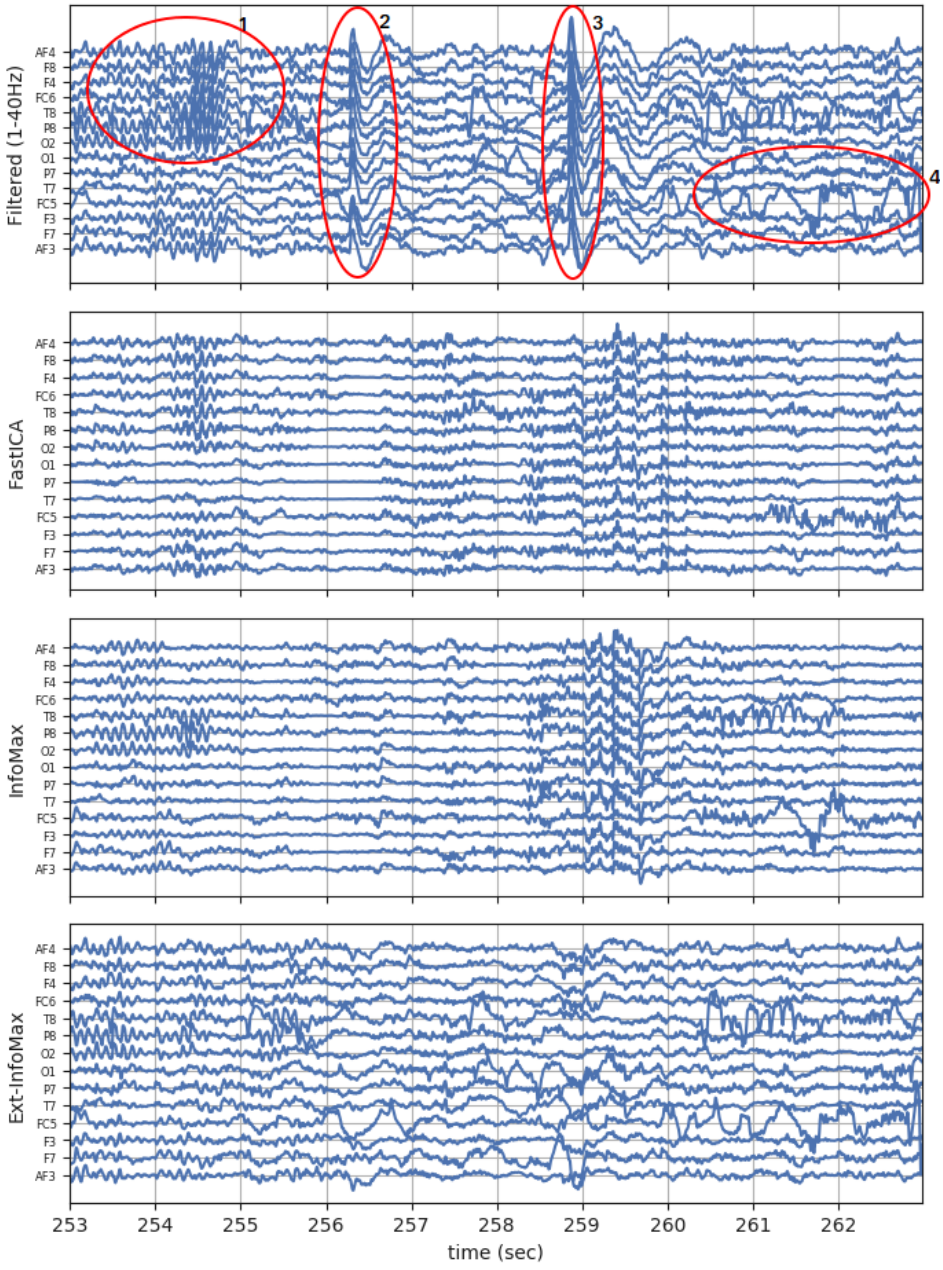


Fig. 4.7 A segment of 10 seconds of 14 channels of EEG signals and corresponding corrected segments by FastICA, InfoMx, Extended-InfoMax. Artifacts identified are indicated in the top figure. The muscular artifact with label-1, motion artifact with label-2 and 3, and blinking artifact with label-4.

signal information is also removed, as signal values are suppressed substantially to zero. Extended-InfoMax is able to remove the peaky and muscular artifacts but slow oscillating artifacts are present in the resulting signal.

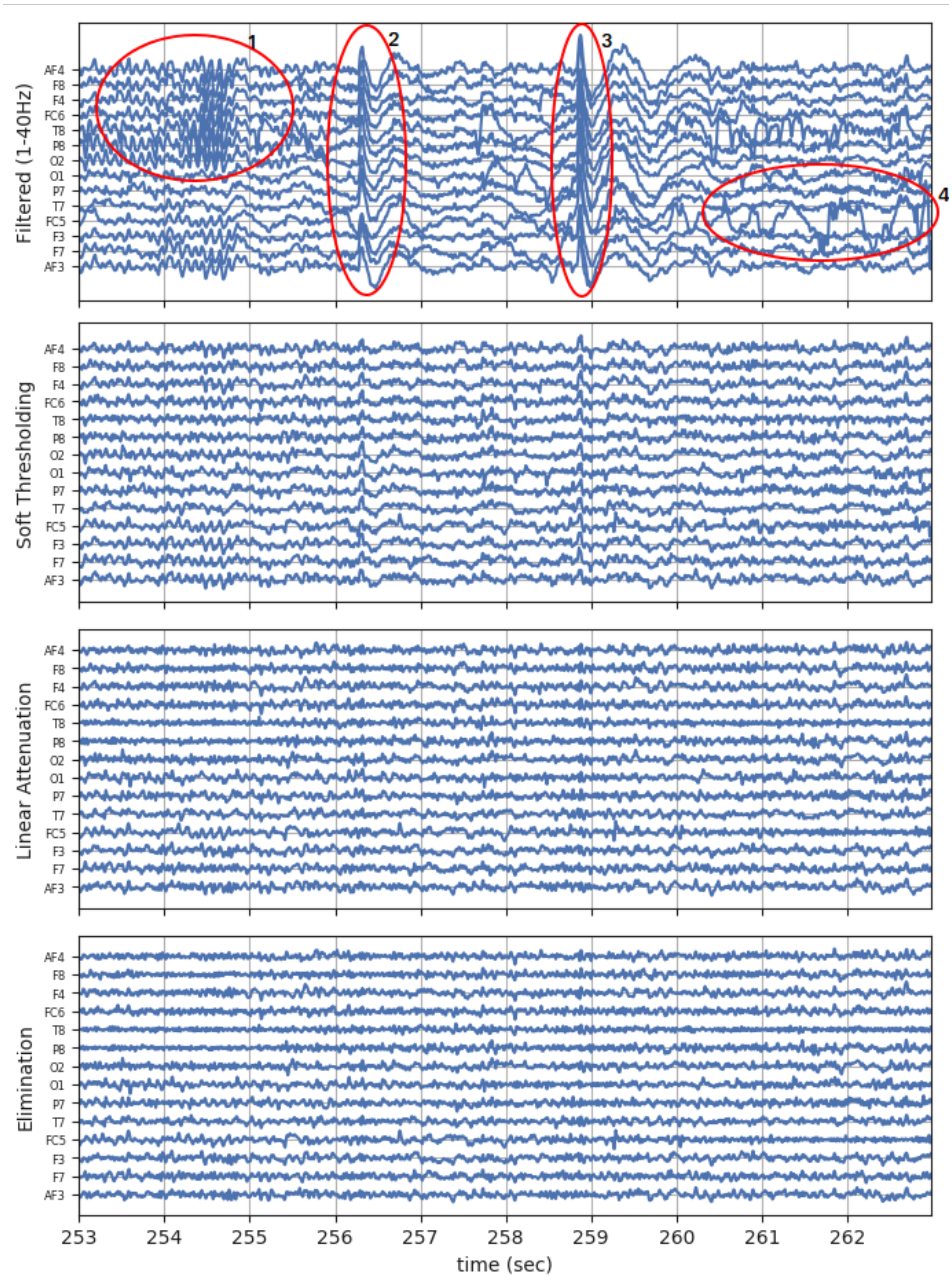


Fig. 4.8 A segment of 10 seconds of 14 channels of EEG signals and corresponding corrected segments by proposed algorithm with $\beta = 0.6$ and $IPR = 50$ for Soft-thresholding, Linear attenuation and Elimination mode of wavelet filtering.

By analysing the resulting signals of the proposed WPD algorithm in Figure 4.8, it can be noticed that for Soft-thresholding all the three kinds of artifacts are substantially suppressed, though a small presence of peaky and muscular artifacts are left. The slow oscillating artifacts are completely removed. On the other hand, as designed, Linear attenuation suppresses the

artifacts even more and Elimination removes all the artifacts. Since the proposed algorithm works on each channel individually, it is interesting to observe that artifacts present in a single channel (FC5) are also removed, which does not happen with ICA, as ICA works on all the channels together to identify the artifactual components. As per design, the effects of this suppression can be tuned. A closer look at the performance of wavelet filtering is shown by a signal channel in Figure 4.9. It can be observed that as designed, the soft threshold is suppressing the signal only where it has higher values, almost preserving the shape of the signal. While linear attenuation and elimination remove the dominant shape of the part. It is worth noticing that at $t = 258$ s, the signal is unaffected by all the modes of wavelet filtering.

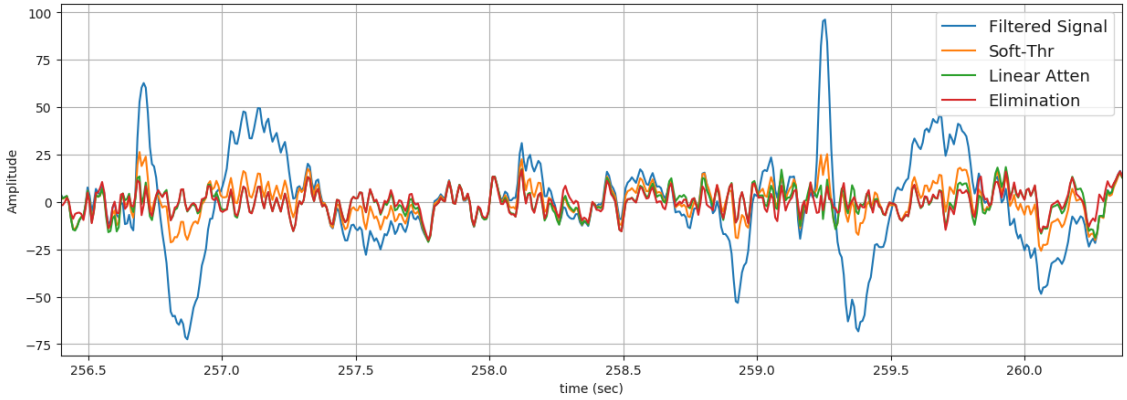


Fig. 4.9 A segment of single channel EEG signal and corrected signal with proposed algorithm for $\beta = 0.6$ and $IPR = 50$ for soft-thresholding, linear attenuation and elimination mode of wavelet filter.

4.5.2 Spectral and amplitude analysis

The power spectral density (PSD) computed with the Welch method for a single filtered channel and the corresponding corrected signals are shown in Figure 4.10. From that figure, it can be observed that the filtered signal has a high density in low frequencies. From the PSD point of view, all the three ICA based approaches are similar at low frequencies, suppressing them and highlighting the 1 Hz and 10 Hz components. On the other hand, our algorithm suppresses low frequencies and highlights also the 22 Hz component beside the 1 Hz and 10 Hz ones. In addition, the spectrum resulting in our algorithms is smoother than that obtained through the ICA-based approach. Again, it can be observed that the amount of suppression increases in the order Soft-thresholding, Linear attenuation, and Elimination. It is interesting to note that, with $IPR = 70$ and Linear attenuation and Elimination mode, lower frequencies are further suppressed even to remove the prominent frequencies of 1 and 10 Hz.

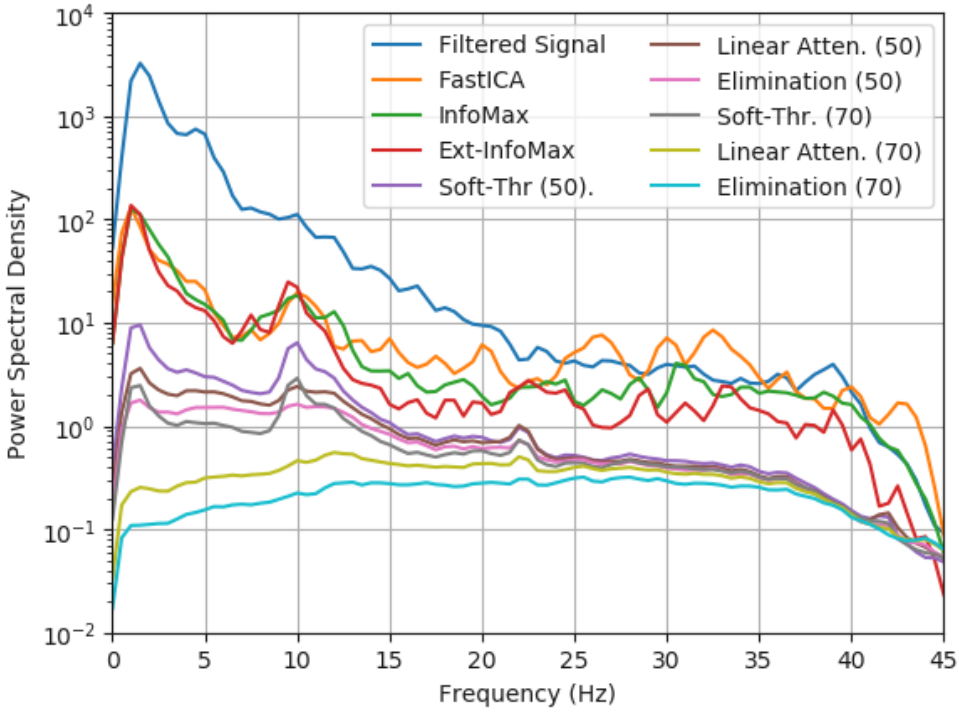
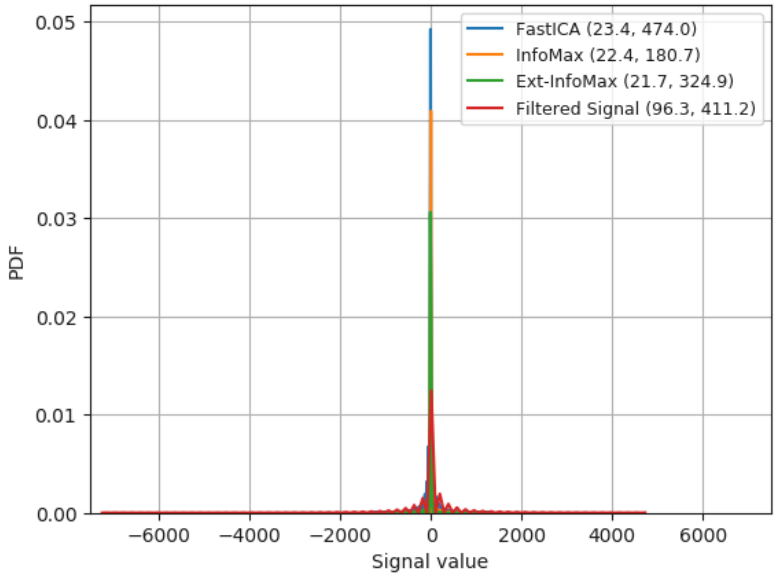


Fig. 4.10 Power spectral density of the filtered and corrected signals. The proposed algorithm *IPR* is indicated in brackets.

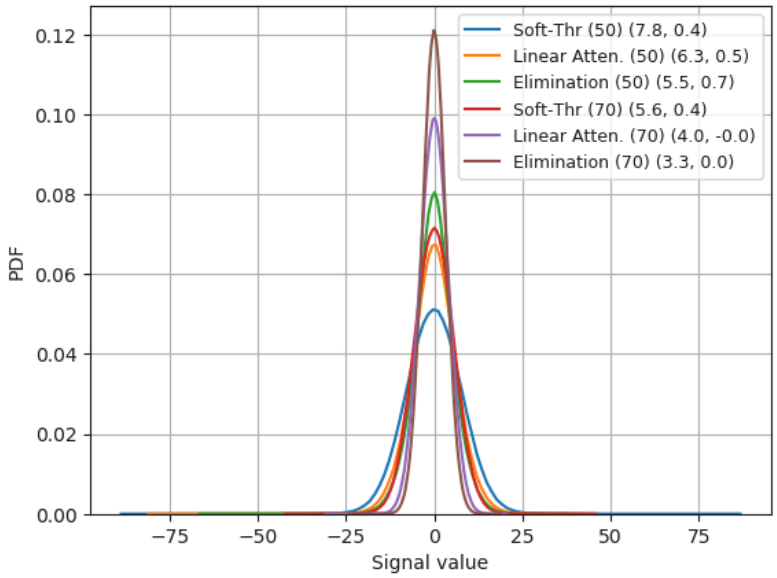
As mentioned, the artifacts cause high variance in the EEG signal. The next approach is to analyze the distribution of the filtered and the corrected signals. Figure 4.11 shows the distribution of the EEG signal values. Figure 4.11a shows the distribution of a filtered signal and the corresponding corrected signal with ICA-based algorithms, and Figure 4.11b shows distribution of signals corrected with our WPD algorithm with different modes and parameters, indicated in the same figure. As shown in the figure, the filtered and the ICA-corrected signals have high SD and kurtosis values (indicated in brackets: SD, kurtosis), strongly suggesting the presence of outliers with high tailed and peaky distribution. Observing the signal corrected with our algorithm, on the other hand, it can be noticed that SD and kurtosis values are drastically lower, SD is under 10 and kurtosis is under 1, for all the algorithm settings. It is worth noticing that kurtosis of Linear attenuation and Elimination for $IPR = 70$ (as indicated in bracket) is reduced to zero.

4.5.3 Correlation of spectral features with target values

As the objective of the artifacts removing algorithm is to improve the performance of the predictive models of auditory tasks, for which EEG signals have been captured, one of the



(a) Filtered EEG signal and corresponding corrected with ICA-based algorithms, FastICA, InfoaMax, Ext-InfoMax



(b) Corrected signals with proposed algorithms, Soft-Thresholding, Linear Attenuation and Elimination, with $IPR = 50$ and 70

Fig. 4.11 Probability distribution of the EEG and corrected signals with standard deviation (SD) of signal and kurtosis of corresponding wavelet coefficients in brackets $(SD, kurtosis)$.

quantitative ways to assess its performance is to measure the improvement in the correlation between input features and target values for predictive modeling. As explained in Section 4.4.1, we have 144 listening segments of 14 channels of EEG, for a participant. Each segment

is labeled with a level of noise and semanticity of audio stimulus played while listening, and the corresponding attention score. For the sole purpose of assessing the algorithm performance, we use the spectral features of the EEG signal in six frequency bands for each channel. The six frequency bands are delta $(0.1-4Hz)$, theta $(4-8Hz)$, alpha $(8-14Hz)$. beta $(14-30Hz)$, low gamma $(30-47Hz)$ and high gamma $(47-64Hz)$, as commonly used in EEG studies. For each band, the sum of the absolute spectral power is computed for each channel of each EEG segment. The features were extracted segment wise as explained in Section 2.7. As there are fourteen channels and six frequency bands, we have eighty-four features ($6 \times 14 = 84$) per segment. For attention score, the target values range from 0 to 100, for semanticity, the target values are 0 and 1, and for noise level the target values are -6 dB, -3 dB, 0 dB, 3 dB, 6 dB and ∞ dB. Since the target values of attention score and noise level are ordinal, we computed Spearman's rank correlation and for semanticity, we computed the point-biserial correlation. Correlation between each spectral feature (out of 84) and target values was computed by varying the θ_α threshold from 1 to 200 with the soft-thresholding mode of wavelet filtering.

The maximum absolute correlation of the eighty-four features and the target value (e.g. attention score, noise level, and semanticity) is plotted against the θ_α threshold in Figure 4.12. In this figure, the dashed lines show the maximum absolute correlation of the eighty-four features when no algorithm is applied and the area above it is shaded with the corresponding color. All the max correlation values have p-value $p < 0.05$ (p-value) except one, at $\theta_\alpha = 2$, with $p = 0.055$ for the attention score. Figure 4.12 suggests that an intermediate value of the threshold, around 50, is good for maximum correlation. The correlation analysis suggests that applying the algorithm increases the chances (with 95% confidence) of at least one feature to have better correlation with the target values, which in principle should improve the performance of predictive tasks.

4.5.4 Performance of predictive tasks

For the predictive modeling tasks as described in Section 4.4.1, the spectral features are extracted as explained in Subsection (4.5.3). For noise level, semanticity and attention score prediction, features are extracted from listening segments only, while for LWR classification, features are obviously extracted from all the segments corresponding to the listening, writing and resting tasks. As explained in Section 4.4.1, LWR, semanticity and noise level predictions are treated as a classification, while attention score prediction is treated as a regression task. For classification, a Support Vector Machine (SVM) classifier, with Radial Basis Function (RBF) kernel of degree 3, is used, while, for regression, Huber regression, with $epsilon = 1.35$ is used. Given our purpose of assessing the performance of artifact removal

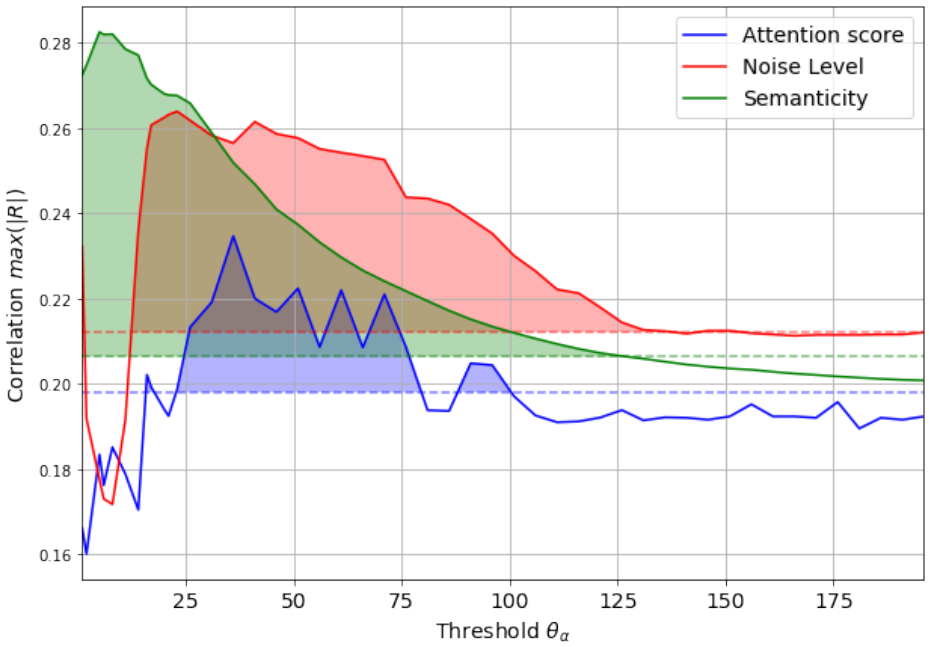


Fig. 4.12 Maximum absolute correlation between spectral features and target value of predictive tasks for different threshold θ_α values with soft-thresholding mode of wavelet filtering.

algorithms, we chose simple and robust classifiers. For performance assessment, Accuracy and Mean Absolute Error (MAE) are computed for classification and regression respectively. Results are computed for the filtered signal, the ICA-based approach and proposed the WPD-based approach, for all the three modes of wavelet filtering explained in Section 4.3.3 along with β as 0.6 and 0.9 and IPR as 50 and 70. The tabulated results (see Table 4.1) are the average performance with 5 fold cross-validation, and the two highest performance values (the testing phase only) among all the methods are highlighted for each task.

It can be observed from Table 4.1 that the ICA-based approach performs better than the baseline model (i.e., the filtered signal), but WPD outperforms the ICA in almost all the cases. For instance, Elimination mode of wavelet filtering with $IPR = 50$ and $\beta = 0.6$ performs better than all the ICA-based approach in all the tasks for testing. Tuning the parameters β and IPR further improves the performance in each task. The highest testing accuracy achieved for the LWR classification is 79.6%, which is much better than for the baseline and for the ICA based approach, which reaches a 75.2% accuracy. Similarly, the highest testing accuracy for semanticity classification and noise level classification is 61.1% and 51.4% respectively, and the lowest testing mean absolute error achieved is 32.746 with Linear attenuation for $IPR = 50$ and $\beta = 0.9$.

Table 4.1 Performance measure with 5 fold cross validation for different tasks; *LWR* classification, Semanticity classification, Noise level-classification, and Attention score prediction. *Tr*- training and *Ts*- testing and MAE is Mean Absolute Error. Two highest performance scores for testing are highlighted.

Method		**LWR** Accuracy		**Sementicity** Accuracy		**Noise Level** Accuracy		**Att. score** MAE	
		Tr	*Ts*	*Tr*	*Ts*	*Tr*	*Ts*	*Tr*	*Ts*
Filtered (baseline)		0.797	0.704	0.753	0.396	0.625	0.493	8.920	37.234
ICA	FastICA	0.793	0.701	0.785	0.424	0.592	0.493	8.153	37.176
	InfoMax	0.828	0.722	0.833	0.458	0.644	0.472	8.509	38.490
	Ext. InfoMax	0.859	0.752	0.898	0.542	0.682	0.444	7.206	36.790
WPD:	Soft-Thr.	0.899	0.766	0.939	0.597	0.802	0.493	7.798	36.893
$IPR = 50$	Lin. Atten.	0.884	0.780	0.892	0.556	0.731	0.486	8.482	38.435
$\beta = 0.6$	Elimination	0.887	0.778	0.892	0.549	0.748	0.507	7.475	33.322
WPD:	Soft-Thr.	0.892	0.782	0.934	0.576	0.804	0.493	8.479	38.899
$IPR = 50$	Lin. Atten.	0.891	0.769	0.910	0.500	0.773	0.493	7.304	**32.746**
$\beta = 0.9$	Elimination	0.883	0.782	0.944	0.583	0.792	**0.514**	8.002	34.891
WPD:	Soft-Thr.	0.894	0.782	0.936	0.583	0.793	0.507	8.235	41.825
$IPR = 70$	Lin. Atten.	0.888	0.773	0.924	0.493	0.781	0.500	7.481	**33.908**
$\beta = 0.6$	Elimination	0.895	**0.796**	0.957	**0.611**	0.818	**0.514**	8.174	35.811
WPD:	Soft-Thr.	0.895	**0.792**	0.934	0.597	0.792	0.500	8.557	39.828
$IPR = 70$	Lin. Atten.	0.892	0.780	0.927	0.514	0.781	0.507	7.701	34.474
$\beta = 0.9$	Elimination	0.894	0.787	0.958	**0.611**	0.809	0.507	7.768	35.542

4.5.5 Effects of β on predictive tasks

As explained in Section 4.3.4, the parameter β is controls the steepness of the curve to select the threshold θ_α. A higher value of β makes curve steeper, which makes the estimation of θ_α lower. The effects of tuning β are quite intuitive for the signal, however, for the predictive tasks, it is not certain, if increasing or decreasing β improves the task's performance. For investigating the effects of β, we computed the performance of predictive tasks, as explained Section 4.5.4 by setting β equal to 0.1, 0.3, 0.6 and 0.9. As discussed in Section 4.5.4, spectral features from segments were computed and the average performance of 5-fold was measured, with *IPR* as 50% and 70%. The results of testing performance are shown in Figure 4.13. From Figure 4.13, it is apparent that for LWR and Semantic classification, applying WPD-based algorithms definitely improves the performance, as all the scores are above the reference line (performance of the baseline model). Furthermore, the performance of these two tasks improves by increasing β from 0.1, which suggests that a higher level of suppression applied to wavelet components highlights the significant predictors to improve the performance.

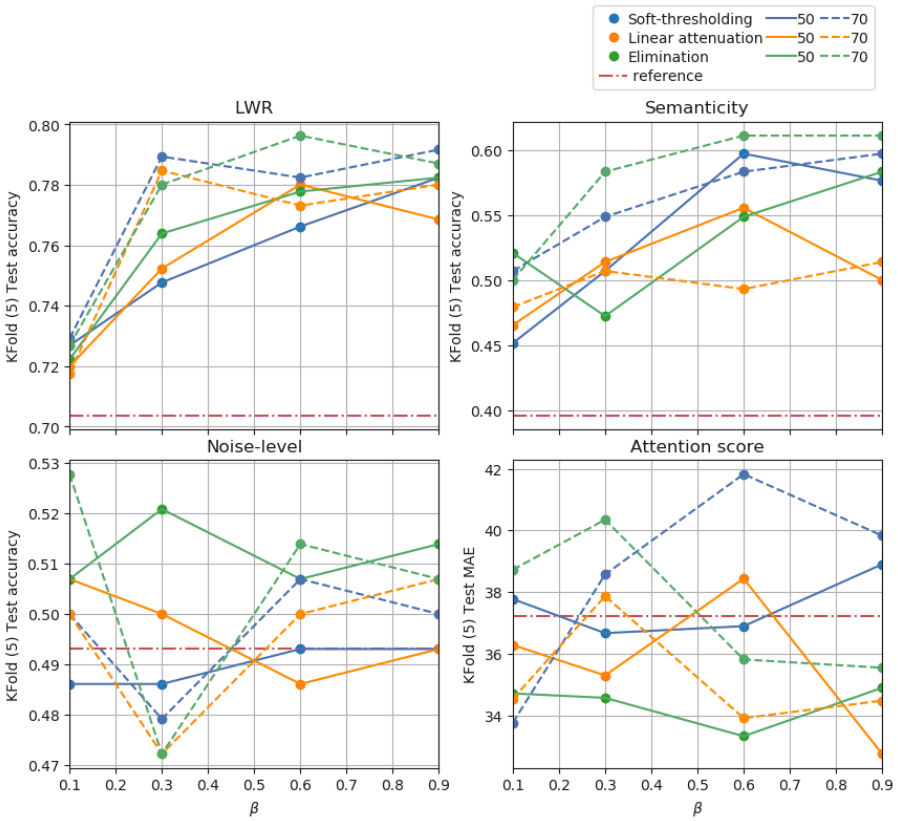


Fig. 4.13 Effect of β on predictive tasks with *IPR* as 50% and 70% for Elimination, Linear Attenuation and Soft-thresholding

For noise level and attention score prediction, however, the performance does not always improve by applying the algorithm. As it can be noticed that performance of noise level decreases at β=0.3, except for *IPR*=50% with elimination, then improves at 0.6 and 0.9 values of β. For attention score prediction, the effect of increasing β with Soft-thresholding mode with *IPR*=70% worsen the performance. On the other hand, Elimination with *IPR*=50% improves the performance and minimum MAE is at β=0.6. The analysis of the effects of β suggests to tune the parameter for the predictive task helps to improve the performance. In addition, it helps to understand the signal values, which are the better predictor for predictive tasks.

In summary, while applying an artifact removal algorithm, it can not be certain whether the algorithm removed any useful information. The state-of-the-art algorithms based on ICA, do not allow to control the suppression or removal of the presumed artifacts. The proposed algorithm based on WPD allows tuning the parameters that control the suppression. In addition, three modes of wavelet filtering provide the choices of assumptions, as to remove or suppress the artifact. We compared the performance of this algorithm with the widely used

ICA-based approach. The proposed algorithm is faster, as it does not require any identify independent components by optimization. Moreover, it does not need several recording channels, as it works on each channel individually. Since the proposed algorithm is based on wavelets, it may be further improved by investigating the selection of particular wavelets or to design one ad-hoc to remove specific kinds of artifact.

Chapter 5

Signal analysis

In this chapter, we will analyse the EEG signals collected during the auditory experiment described in Chapter 2. First, we analyse the spectrum of EEG channels during to different subtasks. Then we analyse the correlation of the spectral bands of each electrode to the corresponding subtasks. We also analyse the Event Related Potential (ERP).All the signal analysis, except for ERP analysis, carried after applying the artifact removal algorithm with soft-thresholding operating mode.

5.1 Spectral analysis

The spectral analysis is one of the most fundamental method in signal processing. The Power Spectral Density (PSD) of recorded EEG signals is estimated by using the Welch method. In Figure 5.1 PSD of a single channel is shown before and after applying the artifact removal algorithm to listening, writing and resting segments. It can be noticed that after applying the artifact removal algorithm, 1 Hz and 10 Hz frequency components are accentuated. Interestingly, the 10 Hz peaks (alpha band) is higher during listening segments and lower during writing segments than resting segment. This distinct pattern is not visible in the spectrum of the raw signal.

Further, to analyse the spectrum of each electrode, we compute the PSD of each channel during listening, writing and resting segments. The data of participant-10 is used and shown in Figure 5.2. The average PSD across all the segments is plotted. For averaging, the median of all the segments is taken to avoid the skewness in the distribution. From Figure 5.2, it can be observed that participant-10 has high alpha activity in all the electrodes while listening, compared to writing and resting. It is interesting to notice that, there is a small peak at 22 Hz in the left hemisphere (left brain - *F7, FC5, T7*, and *P7*) of the brain, but not in the right.

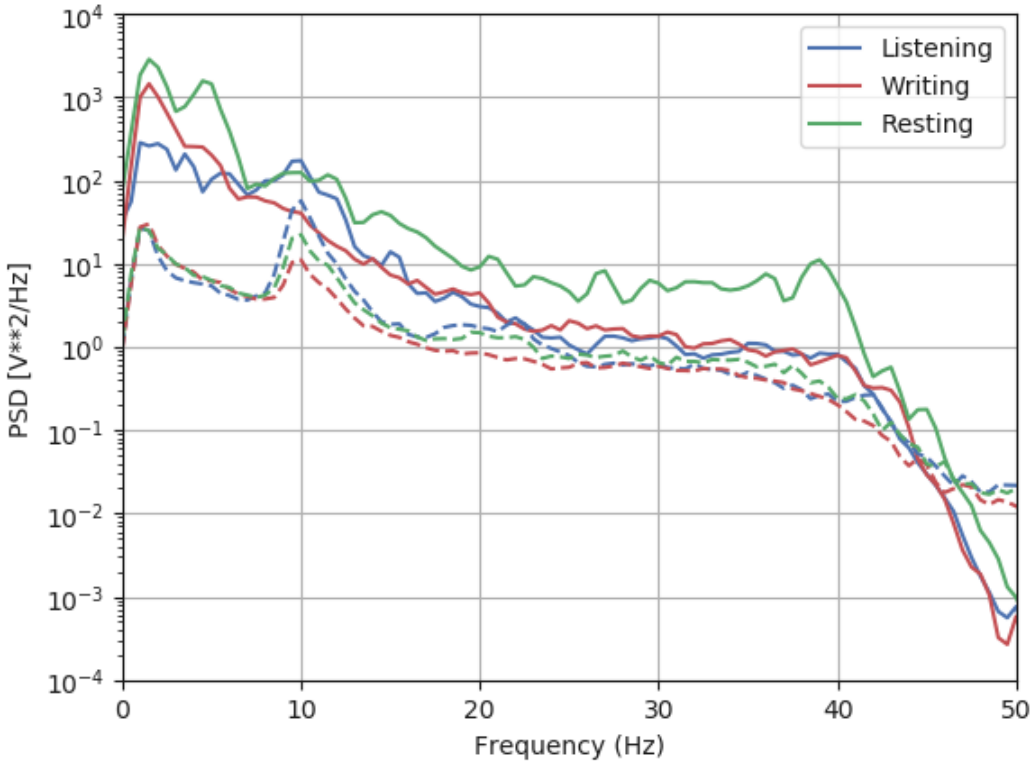


Fig. 5.1 Spectrum of single channel of EEG, before (solid line) and after (dash line) applying artifact removal algorithm.

Surprisingly, participant-1 with the overall highest score, shows the highest alpha activity in the frontal lobe only, compared to temporal and occipital lobe (see Appendix A).

Each participant shows a different spectral pattern (see Appendix A), due to the different folding structure of individual brain [82]. A few participants show alpha activity and others do not. For example Figure 5.3 shows the similar spectral analysis for participant-5 (female, right-handed). Unlike, participant-10, participant-5 shows alpha activities only during the listening task, not during writing and resting task. Spectral plot of all the participants are shown in Appendix A.

Similar spectral analysis is carried out for the noise level and semanticity of stimulus. Figure 5.4 shows the spectral analysis for two extreme cases of noise level; noiseless (SNR=∞ dB) and noisy (SNR=-6 dB) environment. Figure 5.5 shows the analysis for semantic and non-semantic stimulus. For both cases, no significant change in the spectrum has been observed. In general, the alpha activities of participant-10 are higher than participant-5. For quantitative analysis of the spectral response to different activities, a correlation of spectral power in different bands with activities is analysed, which is explained in the next section.

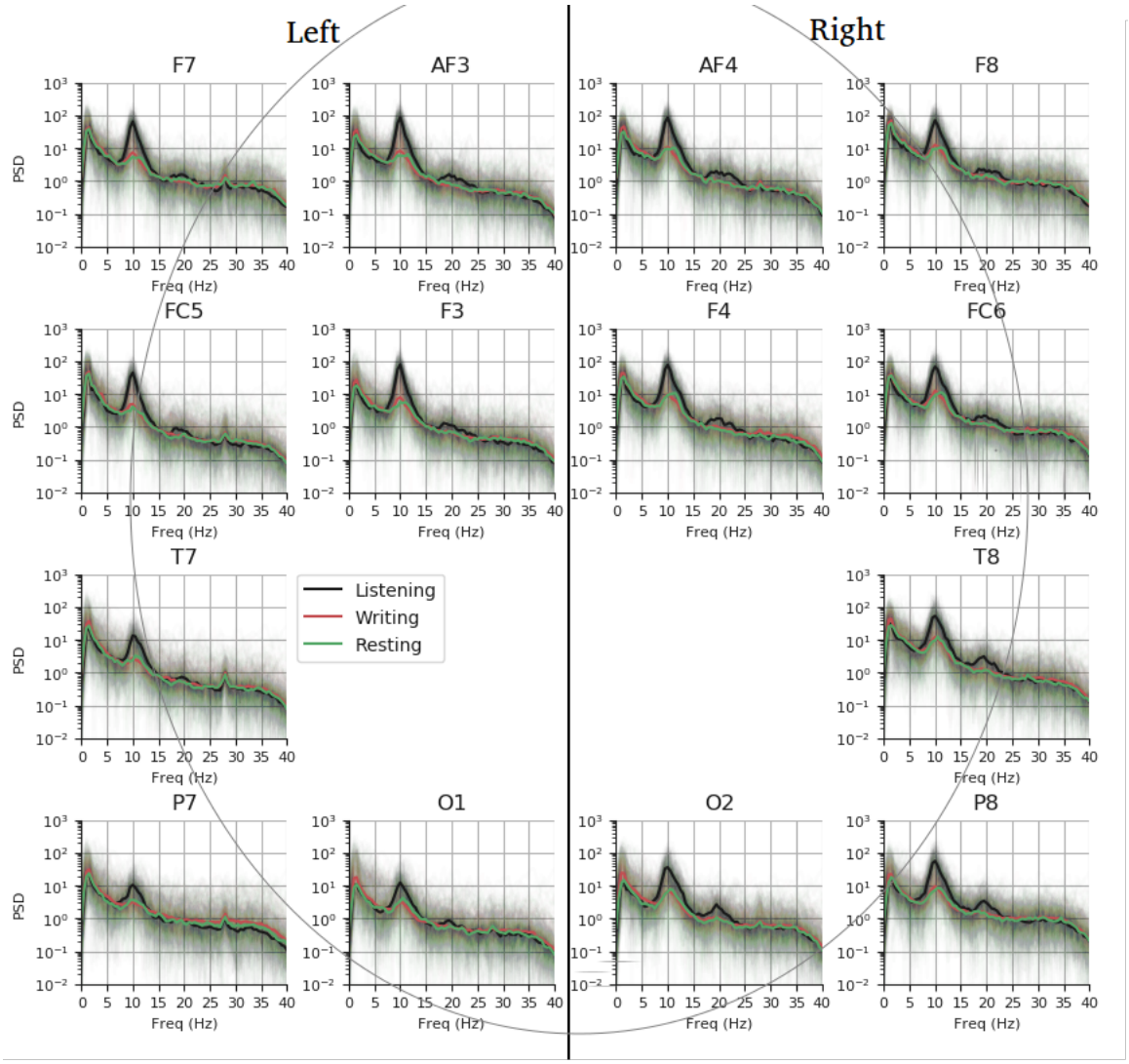


Fig. 5.2 PSD of each electrode of participant-10 (Male, right handed) for Listening, Writing and Resting, arranged in 10-20 system.

5.2 Correlation analysis of EEG

For analysing the correlation of recorded EEG signals with the attention score, noise level, semanticity, and subtasks, all the 14 EEG channels were filtered with a highpass IIR filter with cut-off frequency of 1 Hz and order 5. Artifacts were removed by using a wavelet-based method (explained in Chapter 4). After preprocessing, all signal segments corresponding to the listening, writing, and resting subtasks were extracted for further analysis. Each EEG channel of each segment was firstly decomposed into different frequency bands, namely delta ($0.1 - 4$ Hz), theta ($4 - 8$ Hz), alpha ($8 - 14$ Hz), beta ($14 - 30$ Hz), low gamma ($30 - 47$ Hz) and high gamma ($47 - 64$ Hz). Then, we obtained the sum of the absolute power for

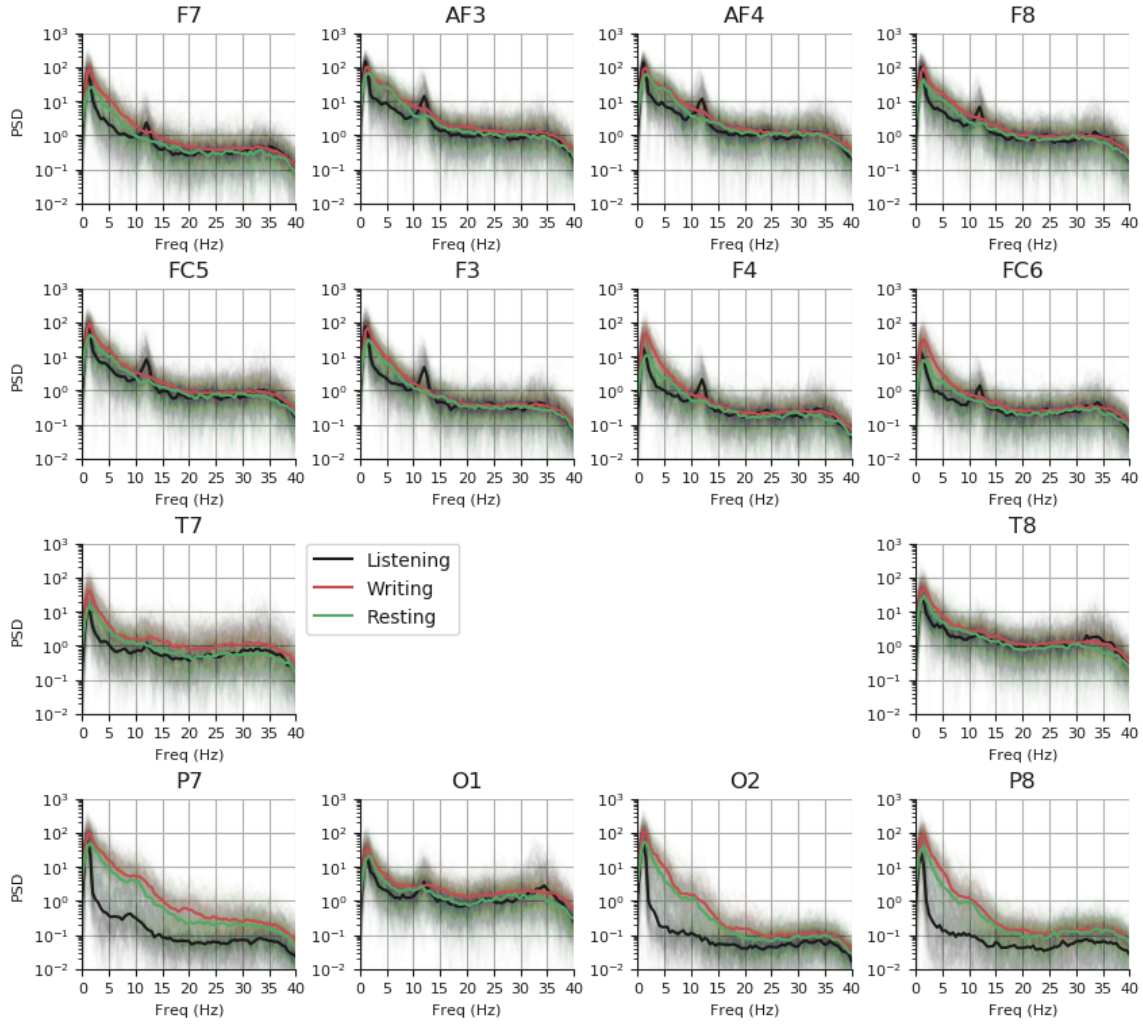
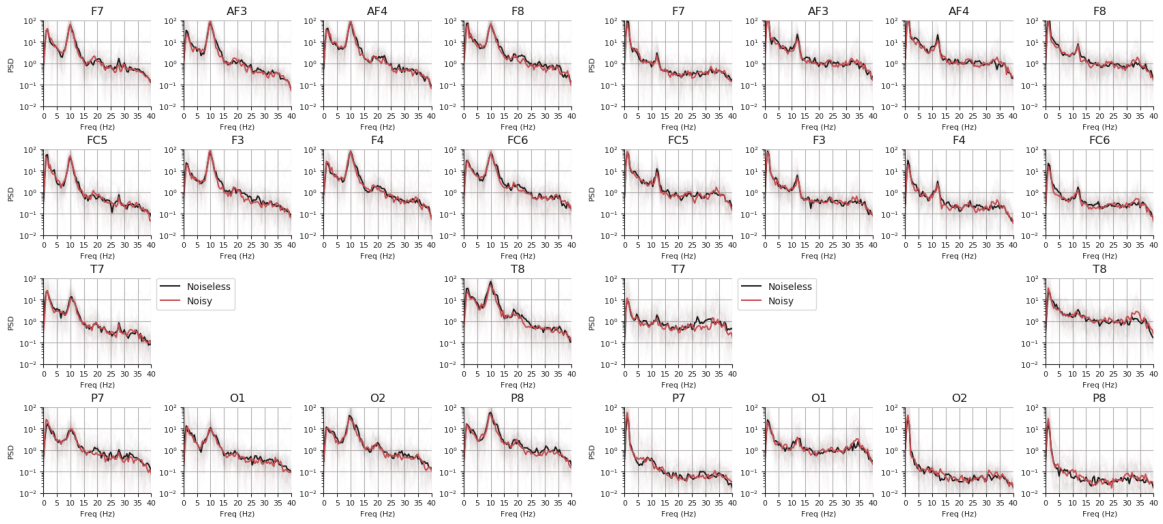


Fig. 5.3 PSD of each electrode of participant-5 (Female, right handed) for Listening, Writing and Resting, arranged in 10-20 system.

each decomposed signal using the Welch method with Hamming windows. The resulting collection of spectral features was therefore calculated as

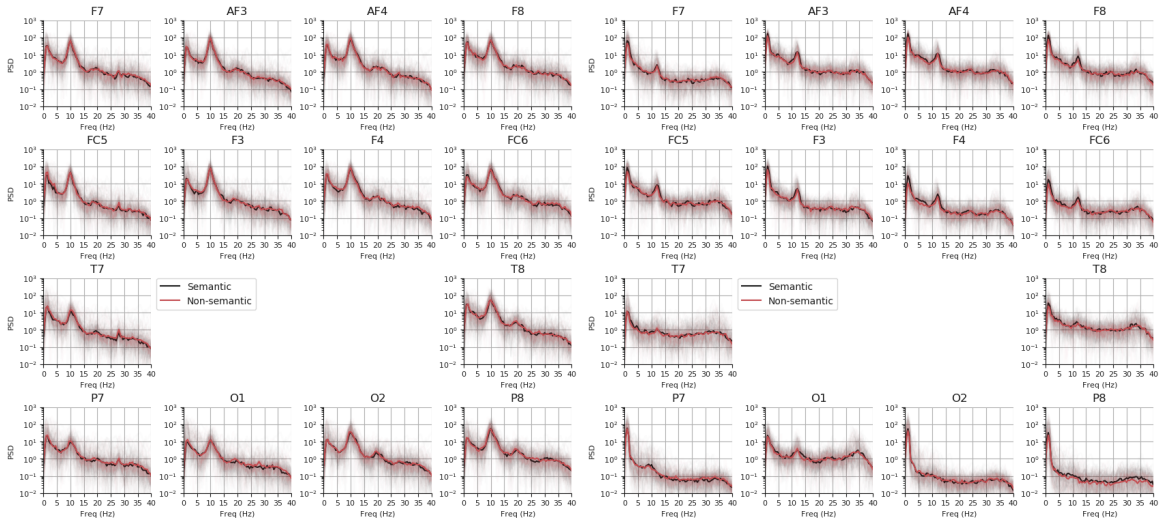
$$F_{i,j,k} = \sum |P_j(E_k^i)|$$

where E_k^i is the i^{th} segment of the k^{th} EEG channel and $P_j(\cdot)$ is the power spectrum of the j^{th} frequency band. As there are 14 channels and 6 different frequency bands, i.e. $k = 1, 2, .. 14$, and $j = 1, 2... 6$, the dimension of the spectral feature vector $F_{i,j,k}$ is $F \in \mathbb{R}^{3N\times6\times14}$, where N is the number of trials for a participant. For N-trials, there are $3N$ total segments. Since attention score, noise level, and semanticity correspond to the listening subtask, the correlation between these factors and spectral power was analysed for listening segments only.



(a) Participant-10 (b) Participant-5

Fig. 5.4 PSD of each electrode for Noiseless and Noisy environment, arranged in 10-20 system.



(a) Participant-10 (b) Participant-5

Fig. 5.5 PSD of each electrode for Semantic and Non-semantic stimulus, arranged in 10-20 system.

As for the correlation between spectral power and the participant's task, all the segments were used.

As the attention score and the noise level are of ordinal data type, the Spearman's rank correlation was computed, whereas, for semanticity and subtask (listening, writing, resting), a point-biserial correlation was computed, as they are of the categorical data

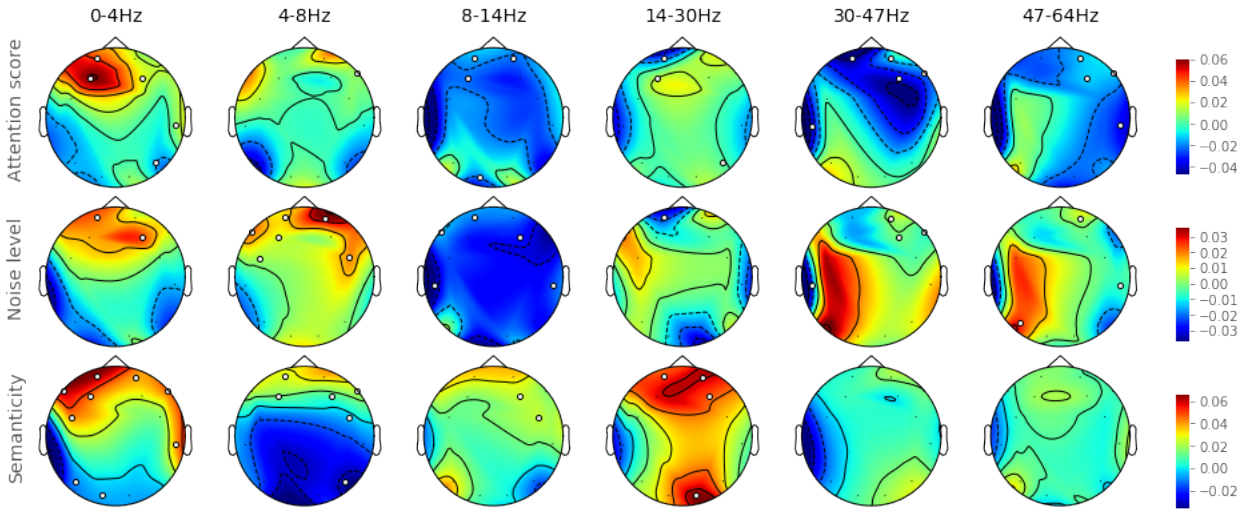


Fig. 5.6 Correlation of spectral power with attention score, noise level and semanticity averaged over all the participants for each spectral band, namely delta ($0.1-4$ Hz), theta ($4-8$ Hz), alpha ($8-14$ Hz), beta ($14-30$ Hz), low gamma ($30-47$ Hz) and high gamma ($47-64$ Hz). The electrodes which correlated significantly ($p < 0.05$) are highlighted with white circular dots.

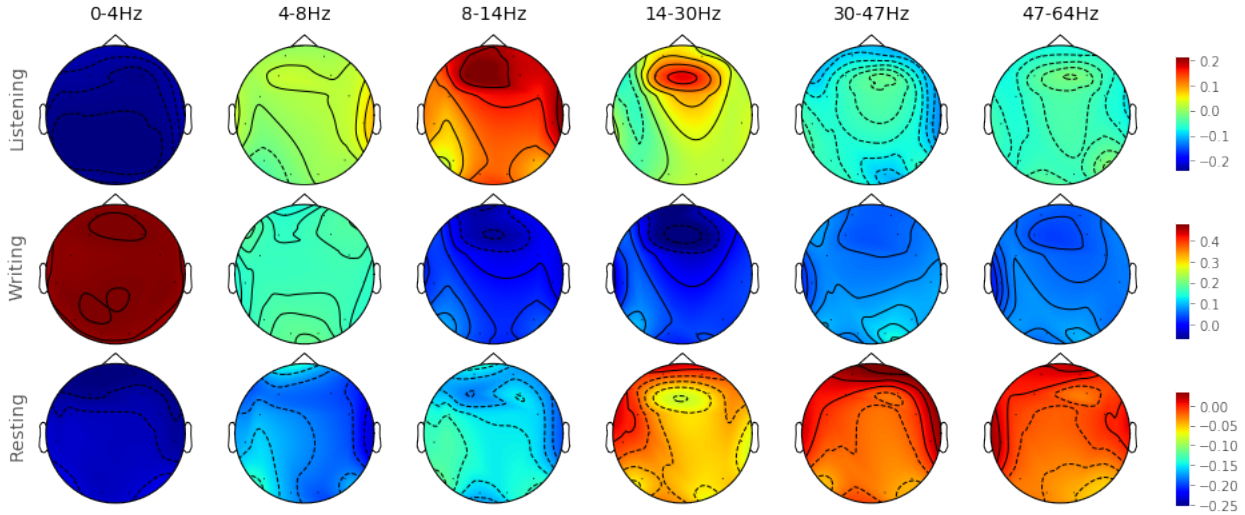


Fig. 5.7 Correlation of spectral power with subtask (listening, writing and resting), averaged over all the participants, for each spectral band. All the electrodes for each band were correlated significantly ($p << 0.01$).

type. The correlation between spectral features and subtask was computed for each subtask individually. A positive correlation indicates an increase in power in the subtask under investigation compared to the other two tasks. The correlation coefficients of each frequency band for each channel were averaged over all the participant with Fisher's Z Method [83]

Table 5.1 The electrodes for which correlation with corresponding activity were significant $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$. The mean correlation $\bar{R}$ along with most negative and positive correlation R^- R^+ are given

	Delta (0.1-4 Hz)				**Theta** (4-8 Hz)				**Alpha** (8-14 Hz)			
	Sensor	R^-	R^+	$\bar{R}$	Sensor	R^-	R^+	$\bar{R}$	Sensor	R^-	R^+	$\bar{R}$
Attention score	AF3**	-1.00	1.00	0.05	F8*	-0.22	0.21	0.01	AF3*	-0.24	0.18	-0.02
	F3**	-0.12	0.27	0.06					F3*	-0.22	0.14	-0.02
	P8*	-0.18	0.25	-0.01					O1**	-0.19	0.22	-0.02
	T8*	-0.17	0.22	0.01					AF4*	-0.21	0.20	-0.02
	F4***	-0.24	0.20	0.02								
Noise level	AF4*	-0.12	0.26	0.02	AF3***	-0.17	0.23	0.01	AF3*	-0.18	0.18	-0.02
	F4*	-0.20	0.18	0.02	F7**	-0.18	0.21	0.02	F7*	-0.20	0.17	-0.01
					F3*	-0.19	0.19	0.01	T7*	-0.27	0.12	-0.04
					FC6*	-0.16	0.16	0.02	T8*	-0.25	0.14	-0.02
					AF4**	-0.20	0.22	0.04	F4*	-0.24	0.16	-0.03
Semanticity	AF3***	-0.12	0.31	0.06	AF3*	-0.2	0.32	0.03	FC6*	-0.21	0.23	0.02
	F7***	-0.15	0.26	0.06	F3*	-0.21	0.22	0.01	F4*	-0.22	0.24	0.02
	F3***	-0.12	0.33	0.05	P8*	-0.25	0.24	-0.03				
	FC5*	-0.10	0.26	0.04	FC6**	-0.20	0.35	-0.0				
	P7*	-0.16	0.19	-0.01	F4*	-0.18	0.31	0.01				
	O1*	-0.13	0.25	-0.01	F8*	-0.19	0.26	0.02				
	T8**	-0.15	0.22	0.04								
	F8**	-0.10	0.35	0.05								
	AF4***	-0.15	0.33	0.04								

	Beta (14-30 Hz)				**Low Gamma** (30-47 Hz)				**High Gamma** (47-64 Hz)			
	Sensor	R^-	R^+	$\bar{R}$	Sensor	R^-	R^+	$\bar{R}$	Sensor	R^-	R^+	$\bar{R}$
Attention score	AF3**	-0.27	0.20	-0.02	AF3*	-0.26	0.19	-0.05	T8**	-0.21	0.21	-0.03
	F3**	-0.24	0.23	0.01	T7*	-0.27	0.17	-0.04	F4*	-0.20	0.25	-0.01
					F4**	-0.34	0.21	-0.04	F8*	-0.20	0.22	-0.01
					F8**	-0.25	0.24	-0.02	AF4**	-0.18	0.22	-0.01
					AF4***	-0.31	0.21	-0.01				
Noise level	AF3*	-0.24	0.11	-0.02	T7*	-0.25	0.19	-0.02	P7*	-0.18	0.18	0.03
					F8*	-0.31	0.12	-0.01	T8**	-0.22	0.21	-0.01
									AF4**	-0.19	0.26	0.01
Semanticity	AF3*	-0.14	0.23	0.05								
	O2**	-0.11	0.23	0.07								
	F4*	-0.12	0.32	0.05								
	AF4**	-0.13	0.30	0.06								

and the corresponding p-values were combined with the Fisher's Method [84], assuming independence [82].

The average correlation coefficients for each channel within each band for the attention score, noise level, and semanticity are shown in Figure 5.6 as a topographic map and the

electrodes that correlate significantly ($p < 0.05$) are highlighted in the same figure. A comprehensive list of the electrodes (sensors) that significantly correlate ($p < 0.05$) is shown in Table 5.1. Similarly, Figure 5.7 shows the averaged correlation coefficients for listening, writing and resting. All the electrodes for listening, writing and resting correlate significantly ($p << 0.01$) in each frequency band.

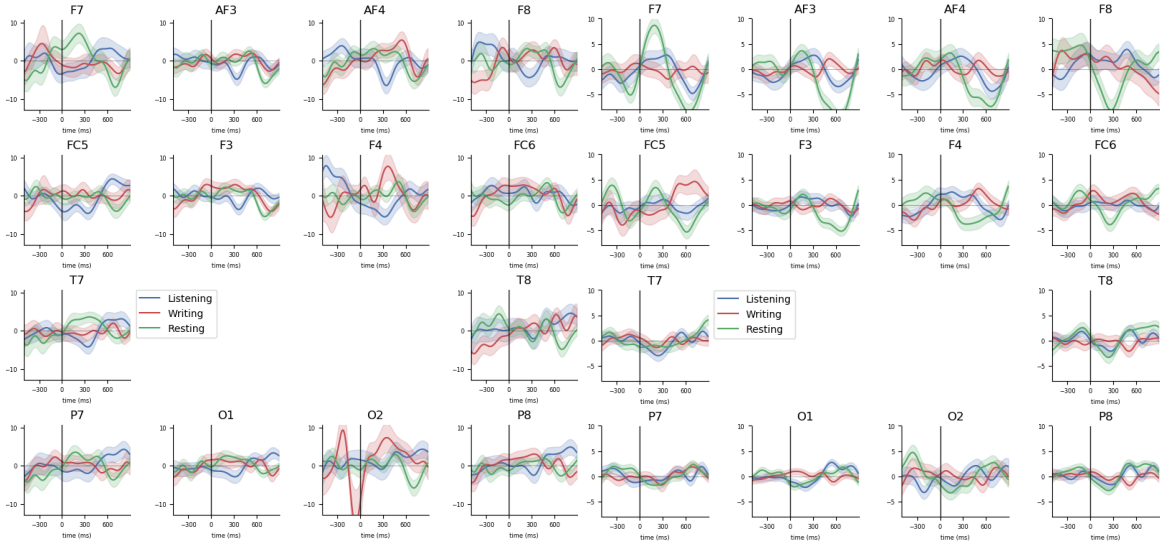
The analysis of the correlation between the attention score and the spectral features reveal interesting properties. For low-frequency bands (delta and theta), the attention score has a significant positive correlation in the frontal lobe, whereas, whereas for high-frequency bands (low and high gamma), it has a significant negative correlation. Indeed, attention has been previously associated with delta activity [85, 86], which is consistent with our results. Interestingly, in the beta band, two frontal lobe channels, namely *AF3* and *F3*, present, respectively, a negative and a positive correlation, as indicated in Table 5.1. By contrast, in the remaining bands, the sign of the correlation is the same for all the channels. Our correlation analysis is also consistent with previous literature reporting an inverse relationship between attention and alpha band [87–90], as shown in Figure 5.6. The analysis of this result indicates that, as SNR increases (i.e. noise decreases), the attention level increases, and subsequently the alpha power decreases. The noise level has a significant positive correlation for low frequencies (delta and theta) around the frontal region, which might indicate some relation between auditory processing and the prefrontal cortex.

Semanticity has the strongest correlation with the delta and theta bands, as indicated in Figure 5.6. It has a significant positive correlation with the frontal and temporal regions and a negative correlation with the occipital region. The positive correlation of the beta band with semanticity is consistent with the study [91], which relates this correlation to active thinking and focus. If the presented stimulus is semantically correct, the power in the beta band increases, as the participant starts thinking actively and focusing on the stimulus. It has also been reported that upper alpha activity is correlated to semantic information processing [92], which explains the slight positive correlation indicated in Figure 5.6.

The spectral power of all the electrodes is significantly affected by the subtask. As mentioned, all the channels present a significant correlation ($p << 0.001$) with listening, writing, and resting. The most prominent effects can be observed in the delta band in Figure 5.7, where all the channels have the most negative correlation with listening and resting and the most positive correlation with writing. In contrast, the correlation reverses for the alpha band during listening and writing. It can be observed that, for resting, high-frequency bands (beta and gamma) are slightly positive, compared to the other two tasks.

5.3 Event Related Potential analysis

The ERP analysis of EEG signals reveals the brain response to any cognitive, motor or sensory event. The millions of neurons in the brain respond to many activities at a time. To isolate the brain response to any specific event, the same event is repeated many times and corresponding brain activities are recorded. Taking an average of all the recorded activities (epochs) with proper time-locked to the event highlights a brain response to an event. For analysing the ERP response, EEG signal if first filtered with a bandpass filter of 1-4 Hz (delta) and then epoch for each channel is extracted with 0.5 seconds prior to the events: listening, writing, resting. For listening, writing, and resting events there are $N = 144$ epochs. Finally, all the epochs are averaged after rejecting the outlier epochs. The outlier epochs are identified by IQR method applied to the average energy of each epoch. We also analysed the ERP response to semantic, non-semantic, noiseless (∞ dB) and noisy (-6 dB) stimulus isolated from listening epochs. We have analysed ERP very briefly, the more intense analysis is reserved for future work.



(a) Participant-10 (b) Participant-16

Fig. 5.8 ERP analysis for Listening, Writing, and Resting.

5.3.1 ERP for subtasks

The ERP response of participant 10 and 16 for subtasks (listening, writing, resting) are shown in Figure 5.8a and 5.8b, arranged in 10-20 system. The standard error of each ERP response is shown with shaded area. The number of epochs before rejecting the outliers for

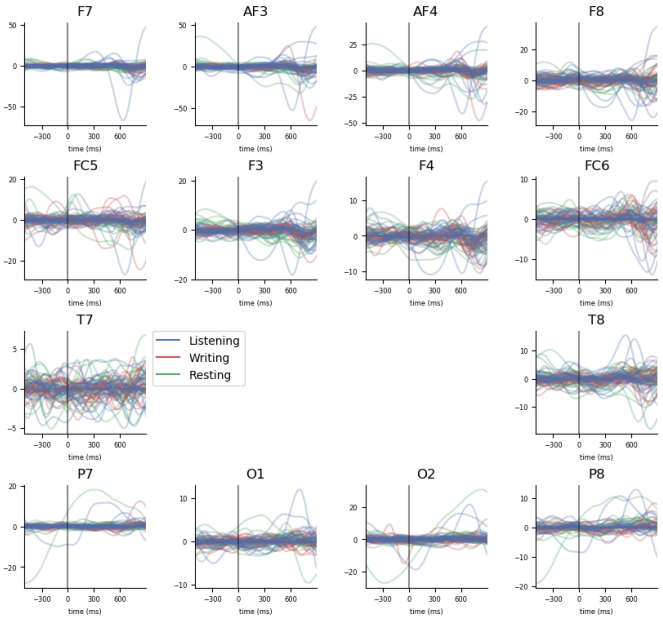
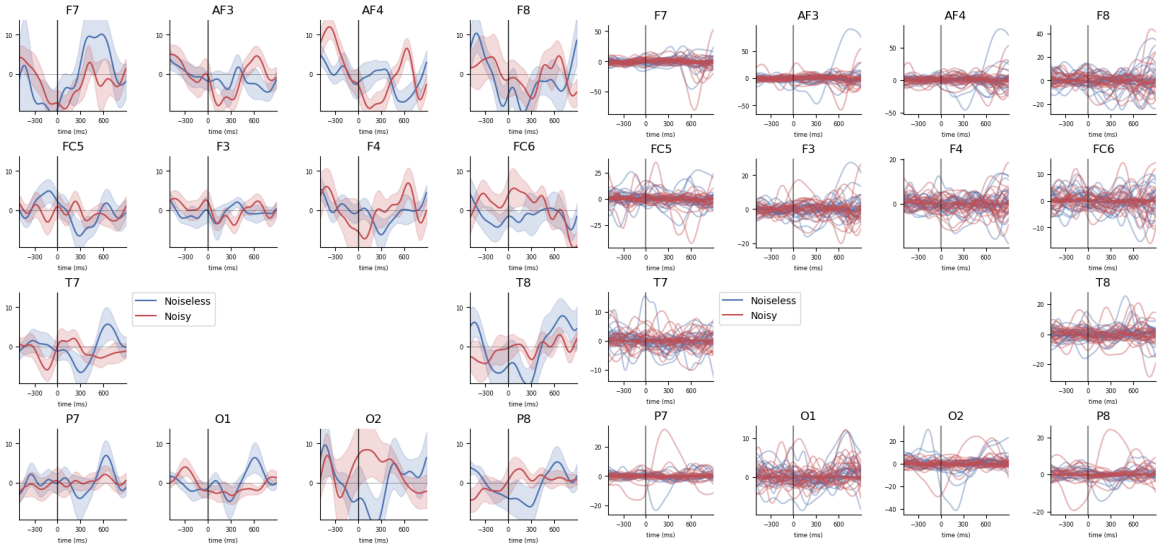


Fig. 5.9 ERP analysis of all the participants for LWR.

subtasks are $N = 144$ (trials) for each participant. From both the Figures, it can be observed that for the listening event, there is a negative peak around 300 ms (usually named as *N2*) more dominating in the frontal lobe and both sides of the temporal lobe in participant 10. Interestingly, a similar negative peak around 300 ms is visible for participant-16, but only in the temporal, occipital, and parietal lobe. For the frontal lobe, a peak is rather at 600 ms. Compared to the listening, ERP for writing event has a strong positive peak around 300 ms in the right hemisphere of participant-10. The resting event has a negative peak in all the electrodes positions around 700 ms for participant-10 and for participant-16, the position of a negative peak is at 300 ms and 600 ms for different electrodes. A reason for a strong ERP response for writing could be due to motor movement. Since each individual behaves differently, ERP analysis is usually done individually. However, for identifying common brain response to the event, averaged ERP of an individual participant are plotted in Figure 5.9. From Figure 5.9, it can be observed that most of the participants have a positive peak ($P3$) around 600 ms for writing and a negative peak for resting in the frontal lobe. Temporal and occipital lobe, however, shows no specific potential response.

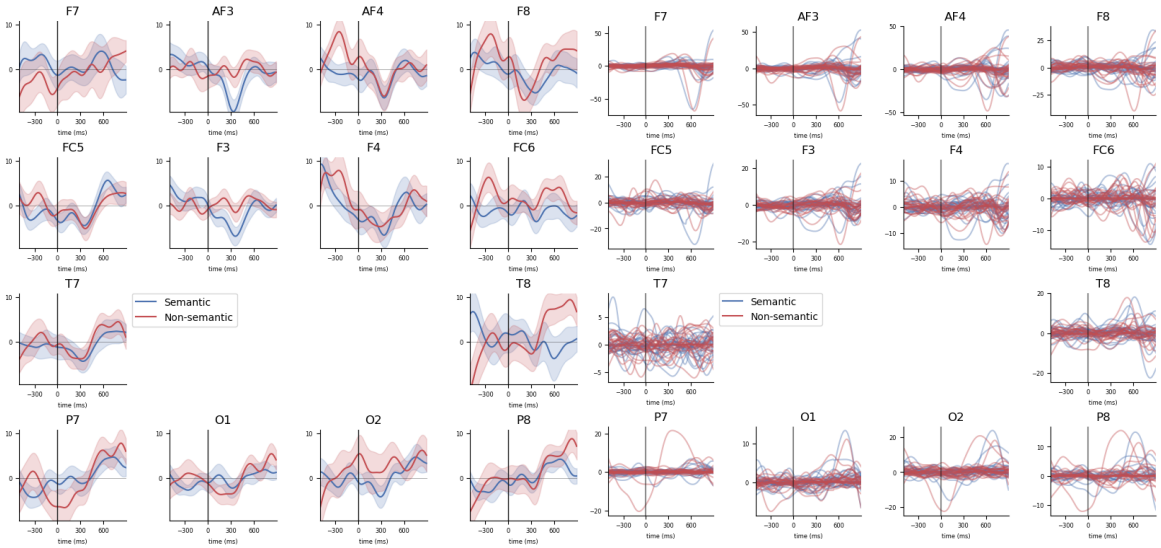
5.3.2 ERP for background noise and semanticity

For ERP analysis of background noise and semanticity of stimulus, epochs were selected from listening segments. The number of epochs for the noiseless and noisy environment were 24 each, for a participant and for semantic and non-semantic 72 each. Figure 5.10



(a) Participant-10 (b) All the participants

Fig. 5.10 ERP analysis for background noise



(a) Participant-10 (b) All the participants

Fig. 5.11 ERP analysis of semanticity

shows the ERP response of participant-10 and the averaged response of all the participants. Interestingly, participant-10 shows a positive peak at 600 ms in temporal and occipital lobe for a noiseless event, at the same location, there is less strong peak is observed for a noisy event, except for prefrontal cortex. Figure 5.10b shows the averaged ERP response of all the participants for noiseless and noisy stimuli. It can be observed that a few participants have a strong negative peak at 600 ms for noisy stimulus in the frontal lobe, A similar analysis of

semanticity is shown in Figure 5.11. Participant-10 doesn't reflect any distinct pattern for semanticity, except for *AF3* and *F3* electrodes, where a negative peak at 300 ms for semantic is more stronger than non-sematic stimulus. For a few participants, there is a positive peak around 500 ms and a negative peak around 600 ms for semantic stimuli, whereas similar peaks are observed for a few other participants for non-semantic stimuli. This opposite behaviour of participants is observed in frontal lobe only.

Chapter 6

Database for scientific use

One of the objectives of the collection of the database is to release it in the public domain for scientific use. Apart from many datasets available for the research, there is a lack of one with auditory attention on natural speech. In this chapter, we describe the database collected from the experiment in details to facilitate the use. We also demonstrate the results of predictive tasks formulated in Chapter 2 using and models such as Support vector machine, Decision tree, and Gradient booster. The database and relevant work will be available at a project homepage - https://phyaat.github.io.

6.1 Related datasets

Processing physiological signals is a well-established research area. In cognitive science, understanding the relationship between physiological signals, such as EEG, GSR, EMG, ECG, and psychological or behavioral processes [20] has opened the door to many applications [21]. Specifically, EEG signals have been used to interface with computers producing systems that are known as BCI. In the early era of BCI, researchers mainly used BCI as a communication tool for disabled people, such as patients with severe motor impairment, by using the BCI speller [93–95], which was then extended for patients in a complete locked-in state [96, 97] or in paralysis [98]. BCI systems have been used also for rehabilitation [99, 100] and as assistive technology [27–30]. Current BCI systems are intended to work not exclusively in health applications, but also in daily-life environments. The ease of recording physiological signals has motivated many researchers to conduct experiments to understand various physiological phenomena of the human body and brain, and psychological processes too. Emotion recognition using physiological signals was investigated in [32–34]. Studies have focused on human emotional state while watching videos [101], listening to music [102], playing games [31, 103, 104], and watching music videos [105]. Assessment of player

states during gaming [106] is another application. These experiments have resulted in several public databases of physiological responses aiming at various activities (e.g., motor imagery movement, sleep analysis, etc.), that have helped the scientific community to collaborate and improve the understanding of human brain and physiology to design effective and robust BCI systems.

Widely used EEG datasets for BCI applications include motor imagery dataset [107–109] , which are supported by other databases for calibration [110], reliability and safety [111]. Other than motor imagery dataset, there are publicly available datasets of physiological responses for sleep [112], epilepsy [113], and emotion [101] analysis.

Focusing on the cognitive abilities of the human brain, datasets of mental arithmetic task [114] and P300 speller of patients with amyotrophic lateral sclerosis [115] are also in the public domain. The datasets of attention-related studies are also in the public domain. A database corresponding to a study conducted with 10 participants and 16 EEG channels for analysing covert and overt visual attention with a P300 speller is available for scientific use [116, 117]. A similar dataset of 8 participants, 60 EEG and 2 electrooculography (EOG) channels for covert shifts of attention has also been released [118].

Specifically for auditory attention with EEG signals, an auditory oddball-type experiment was presented in [119], where the response of participants to oddball sounds inserted in streams of sinusoidal tones was analysed. Using the same auditory oddball paradigm, a database of 21 participants recorded with 60 EEG and 2 EOG channels, named AMUSE, was released [120]. This experiment was based on an auditory speller with spatial hearing cues. In a similar way, a dataset of auditory event-related potential (ERP) speller, conducted with 12 participants was released for a multiclass text spelling application [121]. Another dataset, analyzing the aging effect on auditory-visual attention shift, is also in the public domain [122].

6.2 Database of auditory attention on natural speech

State-of-the-art databases of physiological responses presented in the literature focus on imagery motor movements, visual attention or auditory attention on oddball sinusoidal tones, but to the best of our knowledge lack data about auditory attention on natural speech. This work presents a database focused in this area, with a goal to support future research in the field.

6.2.1 Brief of experiment

The methods and materials used for the experiment are discussed in detail in Chapter 2. Here is brief information of the experiment and collected data. A group of 25 healthy students (4 female, 21 male) participated in the study. Participants were presented to three subtasks namely *listening, writing*, and *resting* with 144 trials. For each trial, a participant is asked to transcribe the audio message presented during listening under different auditory conditions. Three different physiological signals were recorded for an entire duration of the experiment; a 14-channel EEG, GSR, and PPG. In total, 19 streams of signals were recorded corresponding to 14 EEG channels, 2 GSR streams (running average and instant sample) and 3 PPG streams (raw signal, pulse rate, and IBI) at a sampling rate of 128 Hz.

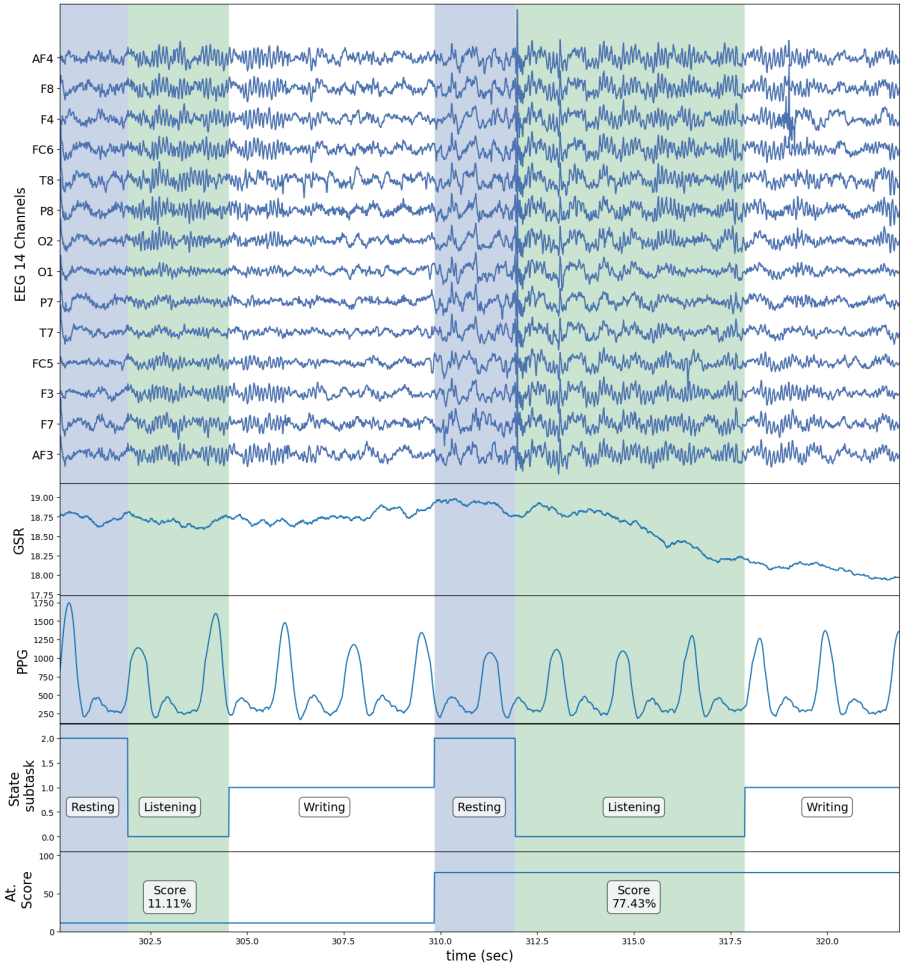


Fig. 6.1 Segments of the 16 signal streams (14 EEG channels, low pass GSR and raw PPG) after preprocessing (highpass filtering and artifacts removal). The intervals corresponding to the subtasks of listening, writing and resting have been highlighted in green, white and blue backgrounds.

6.2.2 Auditory conditions

The auditory conditions presented to the participants during the experiment are based on three factors, namely background noise level (SNR), length of stimulus, and semanticity of stimulus. The different levels of background noise used are -6 dB, -3 dB, 0 dB, 3 dB, 6 dB and ∞ dB (noise free). The length of the stimulus ranges from 3 words to 13 words per stimuli, which were grouped into three categories, L1 (small), L2 (medium), and L3 (long). The category L1 included stimuli of average length 4 words with variation ± 1 word. Similarly, for L2 and L3, average length were 8 and 12. For semanticity, two groups were formed, semantic (labeled as 0) and non-semantic (labeled as 1). The details about the choice and categorization are explained in Chapter 2.

6.2.3 Labeling of physiological responses

All the signal streams were labeled during the experiment by including time, subtask identifier (listening, writing and resting) and experimental condition identifier (level of noise and semanticity of stimulus). The transcription of each audio stimulus for each participant was recorded in a separate file, from which an attention score for each stimulus of all the participant was computed as explained in Section 3.1.1. Finally, a file is created for each participant that includes the attention score, experimental condition identifier, and time respect to each trial. A few segments of preprocessed recorded signals and corresponding labels are shown in Figure 6.1.

6.2.4 Database and file structure

The collected data is recorded in comma separated files (*.csv*) , which is compatible to most of the systems and programming frameworks. The file structure of database is shown in Figure 6.2. The database contains 25 directories, one for each participant, named as *S1* to *S25*. Each directory contains two files; namely *Sx_Signals.csv* and *Sx_Textscores.csv*, where *x* is the participant ID, e.g., *S1* for participant 1. The file named *Signals* includes all the 19 streams of signals at a sampling rate of 128 Hz, along with the subtasks label (*listening, writing,* and *resting* labeled as 0,1 and 2 respectively), and *CaseID* as experimental condition identifier to retrieve the auditory condition and read the relevant attention score from *Textscores* file.

The database also includes a file of the demographic information of the participants, along with their self-rating on English skills. The summary of the database is given in

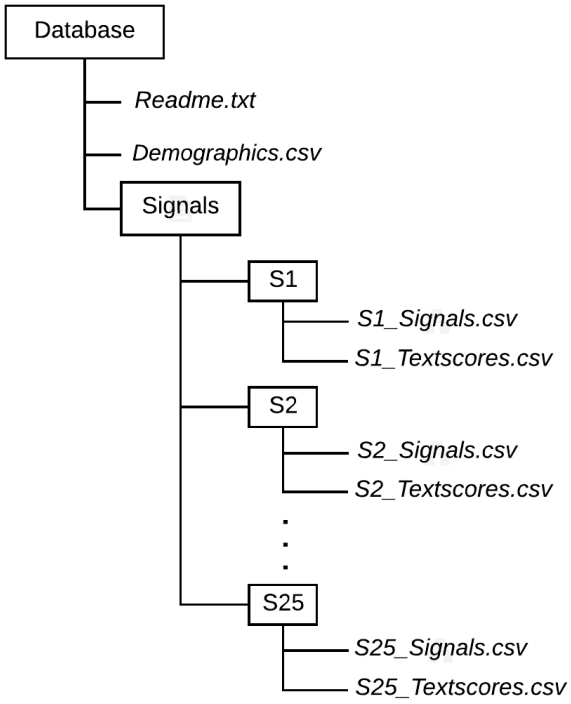


Fig. 6.2 A file structure of database

Table 6.1, and the dataset is available for download from the link[1] (under construction). The demographics, self-rating and overall attention score of all the participants corresponding to their recorded data files are shown in Table 6.2.

6.3 Predictive modeling

In this section, we discuss the performance of the predictive tasks namely LWR, noise level, semanticity, and attention score prediction, as explained in Section 2.7. First, artifact removal algorithm (discussed in Chapter 4) was applied on EEG signal of a participant, with β=0.1 and IPR=50%. Then absolute spectral power for six frequency bands was computed for each channel of EEG, as a feature vector for a segment ($Sg \rightarrow F_r$). The feature extraction was done segment-wise, as explained in Section 5.2. From other five signal streams (2 of GSR and 3 of PPG), we computed mean and standard deviation. For each segment, 94 features were extracted, resulting in a feature vector $F_r \in \mathbb{R}^{84+10}$.

For the LWR classification, there are 3 classes (listening, writing, and resting). For semanticity classification, 2 classes and for noise level prediction, we used 3 classes. Though

[1] https://phyaat.github.io

Table 6.1 Database summary

Physiological signals	EEG, GSR, PPG
EEG channels	14 channels: *AF3, AF4, F3, F4, FC5, FC6, F7, F8, T7, T8, P7, P8, O1, O2*
Total signal streams	19 streams; 14 from EEG, 2 from GSR- instant sample and moving averaged, 3 from PPG - raw signal, beats per minutes, interval between beats (IBI)
Sampling rate	128 Hz
Participants	25 participants: 21 male, 4 female
Number of stimuli per participant	144 randomly selected stimuli, 72 semantic and 72 non-semantic
Independent variables	Noise level, Semanticity, Length of stimulus
Noise levels	6 levels: -6, -3, 0, 3, 6 and ∞ dB SNR
Semanticity levels	2 levels: 0-semantic, 1-non-semantic
Length of stimulus	3 to 13 words per stimulus, grouped into three categories; L1 (small), L2 (medium) and L3 (long)
Average duration of a stimuli	$3(\pm 1.2)$ sec
Average duration of entire recording of a participant	$40(\pm 10)$ mins
Self-rating	Rating scale 1 to 5, for writing, listening, reading and speaking skill of English language

there are six noise levels, we merged them to create three classes only. We merged -3, 0 and 3 dB together and ∞ and 6 dB to one. Resulting three classes are (-6db, 0dB and 6dB).

For prediction modeling, we used Support Vector Machine (SVM), with rbf kernel of degree 3, Decision Tree, Gradient booster with 100 estimators, in addition to Huber regression for attention prediction. Since training and testing are done for an individual participant, Kfold cross-validation with (K=5) is used to evaluate the model performance. The resulting average training and testing performance obtained from 5-fold for a participant are shown in Table 6.3, with a standard deviation of testing performance. As a performance measure; Accuracy and Mean Absolute Error is used for classification and regression respectively.

From Table 6.3, it can be observed that SVM with degree 3 of rbf kernel performs fairly well for classifications tasks (LWR, Semanticity, and Noise level), however for the attention score prediction, SVM is under-fitting (poor performance in training and testing both). Decision Tree, on the other hand, is over-fitting in all the tasks, performing very well on training but not so well on testing. Gradient booster with 100 estimators outperformed SVM for testing in classification and does not worsen the performance for attention score.

Table 6.2 Demographic, self rating, and overall attention score of the participants, corresponding to database files. Self-rating for Rd-Read, Wr-Write, Sp-Speak, and Lt-Listen, at the scale from 1 to 5.

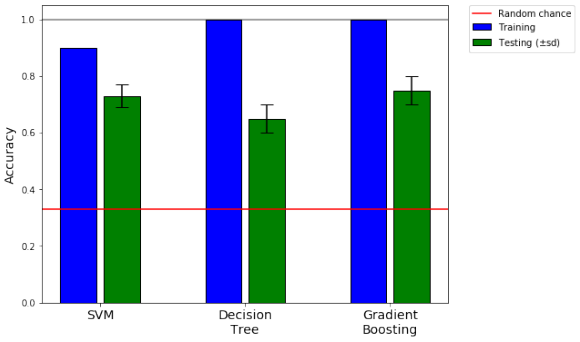
ID	**Age group**	**Sex**	**Nationality**	**First Language**	**Self-rating** Rd/Wr/Sp/Lt	**Overall Att. Score**
S1	26 to 30	Male	Lebanese	Arabic	4 / 4 / 3 / 4	65.86
S2	26 to 30	Male	Indian	Marathi	5 / 3 / 4 / 3	47.77
S3	26 to 30	Male	Indian	Tamil	4 / 4 / 4 / 4	39.70
S6	26 to 30	Female	Indian	Malayalam	5 / 4 / 4 / 4	40.12
S5	21 to 25	Female	Indian	Malayalam	4 / 3 / 3 / 3	35.15
S4	26 to 30	Male	Indian	Malayalam	4 / 4 / 4 / 4	36.15
S7	26 to 30	Male	Indian	Telgu	5 / 5 / 5 / 5	55.61
S8	26 to 30	Male	Algerian	Arabic	4 / 4 / 3 / 3	37.83
S9	26 to 30	Male	Indian	Malayalam	4 / 4 / 3 / 2	49.50
S10	26 to 30	Male	Iranian	Farsi	4 / 4 / 4 / 4	27.13
S11	21 to 25	Male	Lebanese	Arabic	5 / 4 / 4 / 4	31.84
S12	26 to 30	Male	Kazakh	Kazakh	4 / 4 / 4 / 4	35.10
S13	21 to 25	Male	Italian	Italian	4 / 3 / 3 / 3	42.18
S14	21 to 25	Female	Lebanese	Arabic	5 / 4 / 5 / 5	45.54
S15	26 to 30	Male	Iranian	Farsi	5 / 5 / 5 / 5	53.63
S16	31 to 35	Male	Iranian	Farsi	3 / 3 / 2 / 3	29.43
S17	16 to 20	Male	Italian	Italian	4 / 4 / 4 / 4	31.23
S18	26 to 30	Male	Nepali	Maithili	3 / 3 / 4 / 4	44.21
S19	31 to 35	Male	Italian	Italian	3 / 3 / 3 / 3	11.16
S20	26 to 30	Female	Lebanese	Arabic	3 / 2 / 3 / 3	31.63
S21	26 to 30	Male	Pakistani	Urdu	4 / 4 / 4 / 4	23.43
S22	26 to 30	Male	Tunisian	Arabic	5 / 4 / 4 / 4	31.63
S23	26 to 30	Male	moroccan	Arabic	3 / 3 / 2 / 3	14.98
S24	21 to 25	Male	Italian	Italian	5 / 4 / 4 / 4	33.46
S25	21 to 25	Male	Indian	Kannada	5 / 5 / 5 / 5	39.24

From Figure 6.3, it is visible that for all the tasks, the models (except Huber) are performing better than random chance level, that is 0.33 for LWR and noise level classification, 0.5 for semanticity, and 25 for attention score.

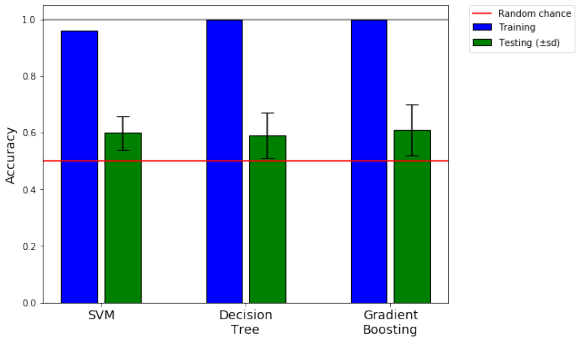
Similarly, the performance of all the four predictive tasks for all the participants is shown in Figure 6.4. For analysing the performance for each participant, only SVM (for classification) and Huber regression (for regression) were used. It can be observed that apart from semanticity classification, performance for all the participant is better than random chance in the classification tasks. For semanticity, the training performance for two participants (10 and 14) is lower than random chance. Similar to Figure 6.3, error bar is shown for the standard deviation of 5-fold cross-validation. The performance of predictive tasks could be improved by employing temporal models such as Dynamic Bayesian Network and Recurrence Neural Network with Long-Short term Memory using temporal features from segments.

Table 6.3 Results of Predictive tasks with 5-fold cross-validation. The average performance for training and testing with different models are listed along with a standard deviation of test performance.

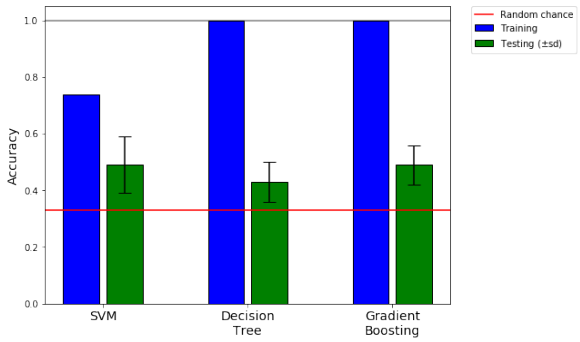
Task	**Model**	**Accuracy/MAE**	
		Training	Testing
LWR	SVM (C=1,'rbf',deg=3)	0.90	0.73 ($\pm$0.04)
	Decision Tree	1.00	0.65 ($\pm$0.05)
	Gradient Boosting (100)	1.00	0.75 ($\pm$0.05)
Semanticity	SVM (C=1,'rbf',deg=3)	0.96	0.60 ($\pm$0.06)
	Decision Tree	1.00	0.59 ($\pm$0.08)
	Gradient Boosting (100)	1.00	0.61 ($\pm$0.09)
Noise level (3 classes)	SVM (C=1,'rbf',deg=3)	0.74	0.49 ($\pm$0.10)
	Decision Tree	1.00	0.43 ($\pm$0.07)
	Gradient Boosting (100)	1.00	0.49 ($\pm$0.07)
Attention score	SVR(C=1,'rbf',deg=3)	14.54	15.14 ($\pm$2.85)
	Huber (ε=1.35, α=0.01)	4.08	25.42 ($\pm$4.10)
	Decision Tree	0.00	21.87 ($\pm$3.89)
	Gradient Boosting (100)	0.68	17.84 ($\pm$2.39)



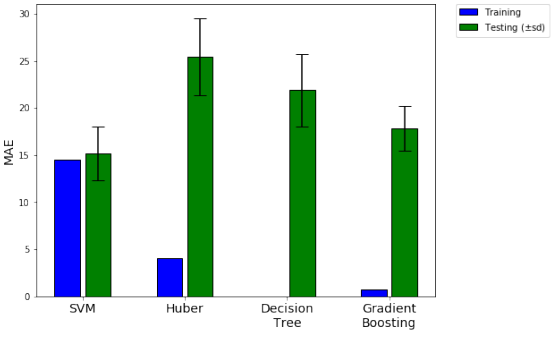
(a) LWR classification



(b) Semanticity classification

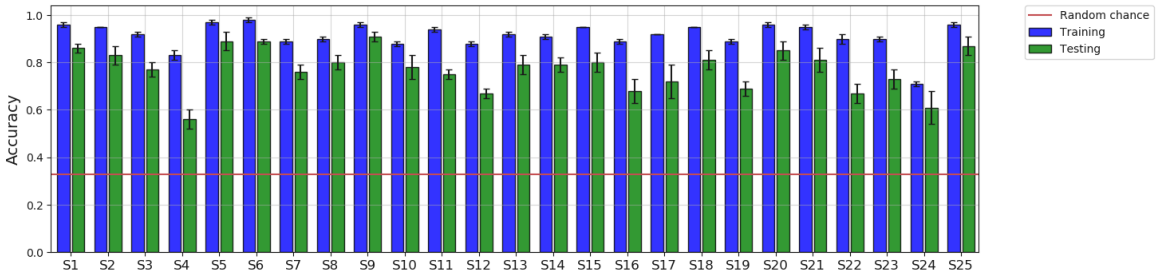


(c) Noise level classification

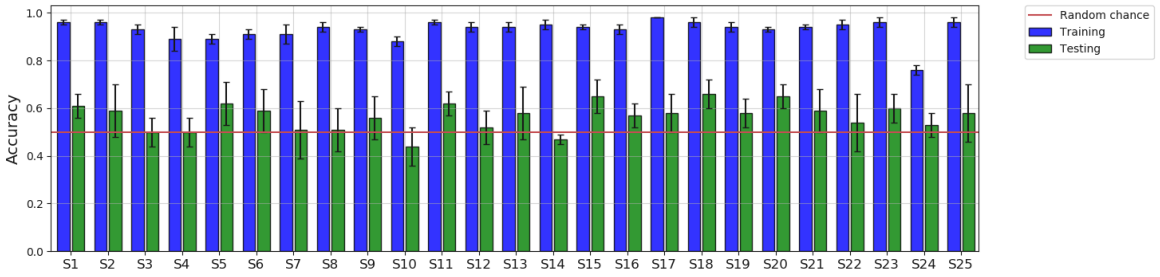


(d) Attention score prediction

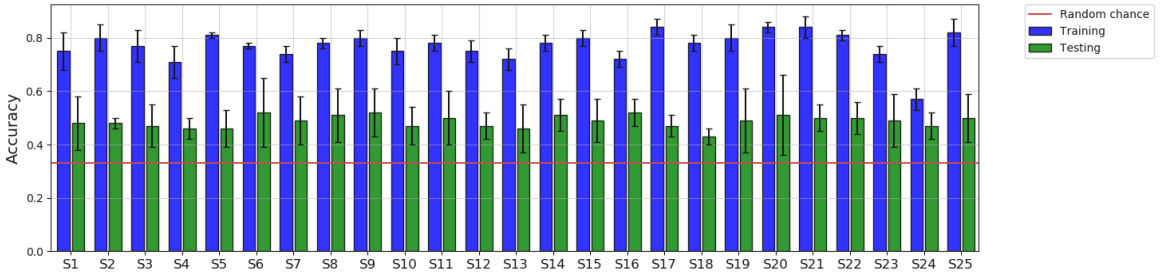
Fig. 6.3 Performance of predictive tasks, using SVM, Decision tree, and Gradient booster.



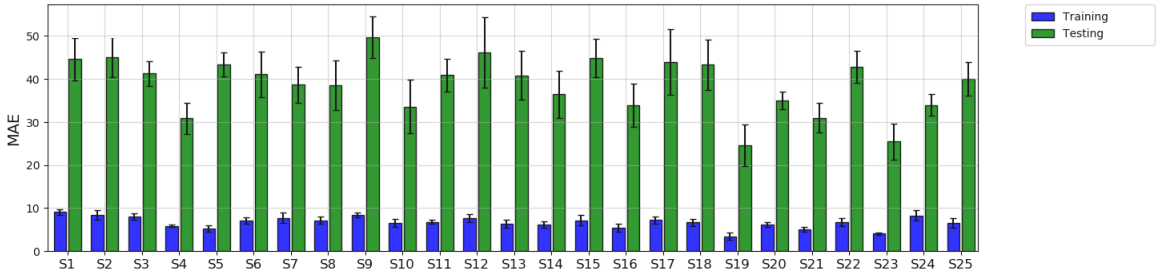
(a) LWR classification



(b) Semanticity classification



(c) Noise level classification



(d) Attention score prediction

Fig. 6.4 Performance of predictive tasks for all the participants using SVM classifier and Huber Regression.